PANTHER User Manual

For PANTHER 9.0

Date: January 7, 2015

Authors:

The PANTHER Team

Contents

1 Welcome to PANTHER System 1
- 1.1 About this document . . . 1
- 1.2 How to cite PANTHER . . . 1
- 1.3 PANTHER help . . . 2
- 1.4 Overview . . . 2

2 PANTHER Home Page 4
- 2.1 Gene List Analysis . . . 4
  - 2.1.1 Available annotation data for the analyses . . . 5
  - 2.1.2 Step 1. Enter IDs or select a list file from your computer or workspace . 6
  - 2.1.3 Step 2. Select organism . . . 8
  - 2.1.4 Step 3.Select Analysis . . . 8
- 2.2 Browse . . . 19
- 2.3 Sequence Search . . . 23
- 2.4 cSNP Scoring . . . 24
  - 2.4.1 Input protein and substitution data . 24
  - 2.4.2 Results of cSNP analysis tool . . . 25
- 2.5 Keyword Search . . . 26

3 PANTHER Pages 29
- 3.1 Subject main page . . . 30
  - 3.1.1 Genes and orthologs . . . 30
  - 3.1.2 Family and HMMs . . . 31
  - 3.1.3 Pathways . . . 32
  - 3.1.4 Ontologies . . . 33
  - 3.1.5 PANTHER Tools . . . 33
  - 3.1.6 Workspace . . . 33
- 3.2 List pages . . . 33
  - 3.2.1 Gene list page . . . 33
  - 3.2.2 Family/subfamily list page . . . 35
  - 3.2.3 Pathway list page . . . 36
  - 3.2.4 Pathway component list page . . . 37
  - 3.2.5 Ontology term list page . . . 39
- 3.3 Information detail pages . . . 40
  - 3.3.1 Gene detail page . . . 40
  - 3.3.2 Family/subfamily detail page . . . 42
  - 3.3.3 Ontology term detail page . . . 44
  - 3.3.4 Pathway description page . . . 45
  - 3.3.5 Pathway molecular class (component) detail page . . . 46
- 3.4 Graph and diagram pages . . . 48
  - 3.4.1 Pie charts . . . 48
  - 3.4.2 Pathway diagram . . . 49
  - 3.4.3 TreeViewer . . . 52

4 PANTHER Basics 56
- 4.1 Download . . . 56
- 4.2 Workspace . . . 57

A Systems requirements 58

Chapter 1

Welcome to PANTHER System

1.1 About this document

This manual provides a step-by-step instruction of how to use all the pages, functions and tools in the PANTHER Classification System website. The manual starts with a detailed description of main features and tools for gene function analysis (Chaptor 2), followed by a description of different types of data pages (Chapter 3) and basic functionalities (Chapter 4).

This manual does not serve as a scientific reference of the PANTHER System. It assumes that you are already familiar with the scientific background of the system. If not, we highly encourage you to read the papers listed in the next section to get familiar with our system.

1.2 How to cite PANTHER

Here is a list of PANTHER publications, based on the subject, that you can cite as references when you publish your results.

- **General PANTHER background** [1]
  Mi H, Muruganujan A, Thomas PD. (2013) PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res.*, 41(Database issue):D377-86

- **PANTHER gene list analysis** [2]
  Mi H, Muruganujan A, Casagrande JT and Thomas PD. (2013) Large-scale gene function analysis with the PANTHER classification system. *Nat Protoc* 8:8, 1551-66

- **PANTHER pathway** [3]
  Mi H, Thomas P. (2009) PANTHER pathway: an ontology-based pathway database coupled with data analysis tools. *Methods Mol Biol.*, 563:123-40.

- **PANTHER software and tools** [4]
  Thomas PD, Kejariwal A, Guo N, Mi H, Campbell MJ, Muruganujan A, Lazareva-Ulitsky B. (2006) Applications for protein sequence-function evolution data: mRNA/protein expression analysis and coding SNP scoring tools. *Nucleic Acids Res.*, 34(Web Server issue):W645-50.

- **PANTHER tree building algorithm (GIGA)** [5]
  Thomas PD. (2010) GIGA: a simple, efficient algorithm for gene tree inference in the genomic age. *BMC Bioinformatics.*, 11:312.

- **Gene Ontology annotation data** [6]
  The Gene Ontology Consortium. (2013) Gene Ontology Annotations and Resources. *Nucleic Acids Res.*, 41(Database Issue):D530-35

1.3 PANTHER help

This manual is not the only help you can get to learn how to use the PANTHER System. You can get additional help and support through the following:

- On most of our webpages, there are links to help tips that provide detailed information about the page, and how to use it. Simply click the icon to get to the Tips page.
- You can always contact us through our feedback email at feedback@pantherdb.org.

1.4 Overview

OK, here we go. The PANTHER home page (Figure 1.1) provides an overview of the PANTHER functionality. The top menu bar (item 1-5 in Figure 1.1) contains links to individual subject main pages with links within that subject. The main page (item 6 in Figure 1.1) contains commonly used tools for users to search, navigate and analyze data. The left panel (item 9-13 in Figure 1.1) contains quick links to the keyword search and other PANTHER information.

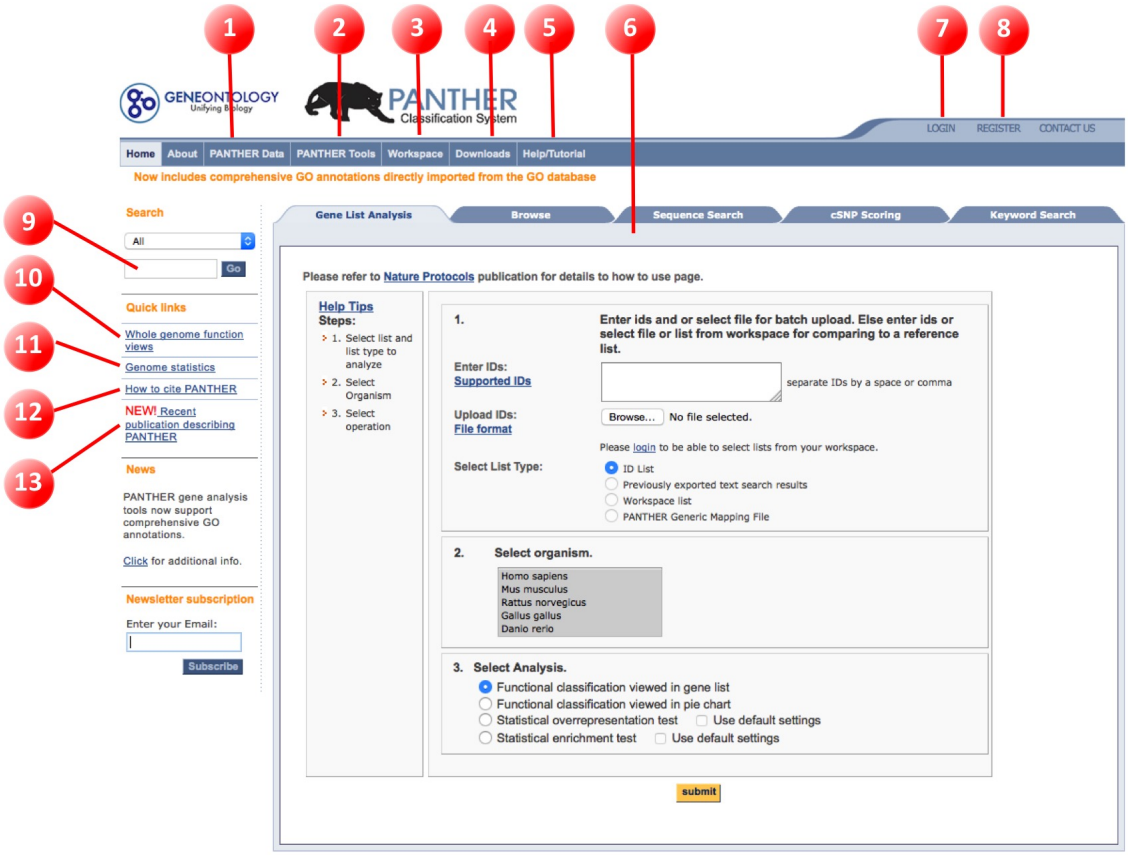


Figure 1.1: *The PANTHER home page*

Below is an overview of the PANTHER home page (the numbers correspond to those in Figure 1.1).

1. **PANTHER Data** – Provide links to different types of PANTHER data (see Section 3.1.1 - 3.1.4 in Chapter 3).

2. **PANTHER Tools** – Links to all the PANTHER tools page (Section 3.1.5. See details in Chapter 3).
3. **Workspace** – Links to the workspace available to the registered users (Section 3.1.6). Registration is free.
4. **Downloads** – Links to the PANTHER FTP site (Section 4.1).
5. **Help/Tutorial** – You can find more help tips and tutorial materials at this link.
6. **Main page** – The main page provide easy access to the most commonly used tools in PANTHER. You can click on the top folder menus to switch among different tools. Details are described in Chaptor 2.
7. **Login** – Click here to login to the website and use the workspace. PANTHER website does not require a login, but the workspace is only available to registered users.
8. **Registration** – It is not required to register to use the PANTHER website. Registration allows you to create a workspace that you can store your analysis data on our site, and it is free.
9. **Quick keyword search** – Perform simple keywords search for genes, proteins, families, ontology and pathway terms.
10. **Whole genome function views** – Links to the whole genome pie chart (Section 3.4.1).
11. **Genome statistics** – Links to page with statistics of 48 genomes in the PANTHER library.
12. **Publications** – This section lists some of the PANTHER publications and information about citing PANTHER.
13. **Recent publications that cite PANTHER** – This section provides a list of selected publications that use PANTHER in their data analysis.

Chapter 2

PANTHER Home Page

PANTHER home page provides direct links to the most commonly used PANTHER tools and features. By clicking the folder tabs at the top, users can switch between different tools.

2.1 Gene List Analysis

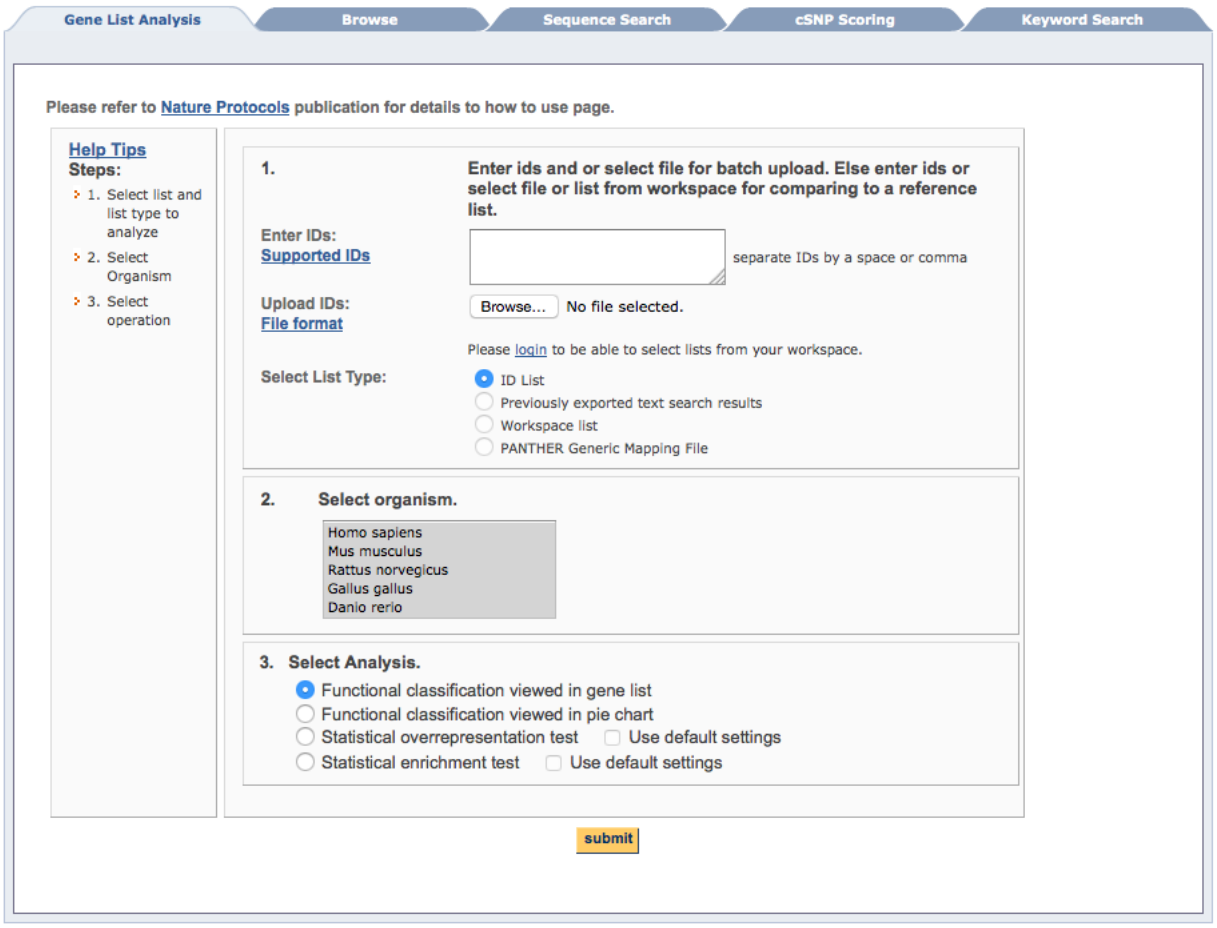


Figure 2.1: *Gene list analysis page.*

The first tab provides a set of tools to analyze a list of genes or proteins (Figure 2.1). This is also the home page when you come to PANTHER. On this page, you can

- Upload a list of genes/proteins, and get functional classification with Gene Ontology, PANTHER Protein Class, and PANTHER Pathway. The results can be displayed as either a gene list page or a pie chart.
- Upload up to four lists of genes/proteins, and run the overrepresentation test to compare them to a reference list to statistically determine over- or under-representation of the ontology categories you select. Binomial distribution test is used in this test.
- Upload a list of genes/proteins, together with values (gene expression, protein expression, etc.), and run the enrichment test to evaluate likelihood of the numerical values of genes associated with a particular ontology term were drawn randomly from the overall distribution of values. The Mann-Whitney U Test (Wilcoxon Rank-Sum Test) is used in this test.

The details of using these tools are described in the following Nature Protocol paper [2]

Mi H, Muruganujan A, Casagrande JT and Thomas PD. (2013) Large-scale gene function analysis with the PANTHER classification system. *Nat Protoc* 8:8, 1551-66.

Since the publication of the above paper, some significant improvements have been made, especially the support of the entire comprehensive GO annotations in the analysis. You can follow the step-by-step Help Tips on the page to run the analysis as below.

2.1.1 Available annotation data for the analyses

- **PANTHER Pathway**
- **PANTHER GO-slim Molecular Function**
- **PANTHER GO-slim Biological Process**
- **PANTHER GO-slim Cellular Component**
- **PANTHER Protein Class**
- **GO molecular function experimental only**- GO molecular function annotations based on experimental evidence published in scientific literatures. These annotations are usually manually curated by GO curators
- **GO biological process experimental only** - GO biological process annotations based on experimental evidence published in scientific literatures. These annotations are usually manually curated by GO curators.
- **GO cellular component experimental only** - GO cellular component annotations based on experimental evidence published in scientific literatures. These annotations are usually manually curated by GO curators.
- **GO molecular function complete** - Complete GO molecular function annotations including both manually curated (above) and electronic annotations. Electronic annotations are generated by computer algorithm based on sequence similarity, and are usually not reviewed by curators. They are less reliable.
- **GO biological process complete** - Complete GO biological process annotations including both manually curated (above) and electronic annotations. Electronic annotations are generated by computer algorithm based on sequence similarity, and are usually not reviewed by curators. They are less reliable.

- **GO cellular component complete** - Complete GO cellular component annotations including both manually curated (above) and electronic annotations. Electronic annotations are generated by computer algorithm based on sequence similarity, and are usually not reviewed by curators. They are less reliable.

2.1.2 Step 1. Enter IDs or select a list file from your computer or workspace

You can input the gene or protein identifiers in three ways.

1. Type into the Enter IDs box. You can also copy/paste multiple IDs to the box. Separate IDs by space or comma. Please refer to the Supported IDs section for the types of IDs supported in PANTHER.
2. Upload a list file from your computer. There are three types of list that are supported by the tools. Please visit File Format link for details of the list types
   (a) ID List
   (b) Previously exported text search results
   (c) PANTHER Generic Mapping File
3. Choose from your Workspace (Figure 2.2). If you registered and have a Workspace, click on the login link and you can upload lists that saved in the Workspace directly for analysis. For details, please visit section 4.2.

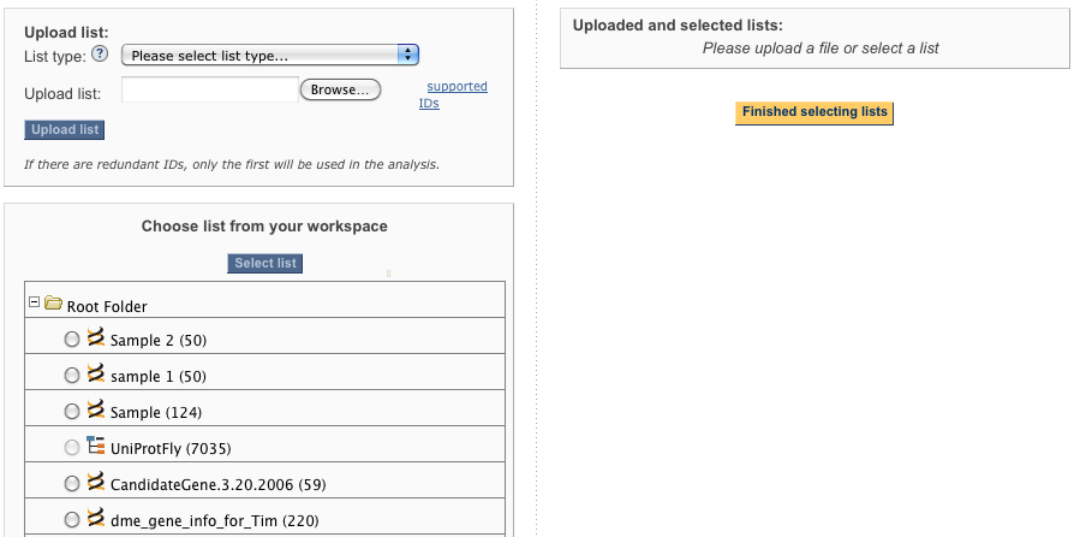


Figure 2.2: *Choose a list from the workspace.*

If you use 2 or 3 to input the IDs, make sure to select the list type that corresponds to your list.

Supported IDs
Not all IDs are supported by the PANTHER. We are working hard to improve this, but at the moment, only the following IDs are supported (You can also click the "supported IDs" link to get the following list). Please note that the system is not able to do partial match so you need to input the full IDs.

- Ensembl: Ensembl gene identifier. Example: ENSG00000126243

- Ensembl_PRO: Ensembl protein identifier. Example: ENSP00000337383
- Ensembl_TRS: Ensembl transcript identifier. Example: ENST00000391828
- Gene ID: EntrezGene IDs. examples include, GeneID:10203, 10203 (for Entrez gene GeneID:10203)
- Gene symbol: for example, CALCA
- GI: NCBI GI numbers. Example: 16033597
- HGNC: HUGO Gene Nomenclature ids. Example: HGNC:16673
- IPI: International Protein Index ids. Example: IPI00740702
- UniGene: NCBI UniGene ids. Examples: Hs.654587, At.36040
- UniProtKB:UniProt accession. Example: O80536
- UniProtKB-ID: UniProt ID. Example: AGAP3_HUMAN

PANTHER Generic Mapping File list type: If your id type is not listed in the above supported ID list, select this list type, and you can analyze IDs that re not found the PANTHER system, by uploading user-generated data containing mappings between any arbitrary identifier and its corresponding PANTHER ID (see below for details about mapping). You can map your sequences to PANTHER, by downloading the PANTHER HMM scoring tools (available at the downloads section on the PANTHER site) and scoring your sequences against the PANTHER HMM library. For sequences that do not match a PANTHER HMM, please assign a PANTHER ID of NOHIT. This is very important, since this information will be used to assign sequences to the "unclassified" category. This will ensure that the statistics can be calculated correctly.

File format

The tool only accepts simple text (.txt) in the following format.

- If your file contains a list of IDs, separate the IDs by *tab, carriage return, line return, comma, pipe* or *space*. if you are uploading a list of IDs with values, the file should contain 2 columns, the first column contains the IDs and the second column contains the values.
- If your file contains previously exported genelist (see section 3.2.1 about how to export a gene list), use the tab-delimited file resulting from the export feature. The system searches on the Gene ID column.
- If you are uploading a **PANTHER Generic Mapping File**, the file must be tab-delimited and contain the following columns:
  - The first column can contain any arbitrary ID (ex: a probe ID), but must be unique, since this allows the user to uniquely specify each record in the dataset, so they can track the identifier on the PANTHER website.
  - The second column should be corresponding PANTHER ID (ex: PTHR10078 or PTHR10078:SF1), and is used to look up the molecular function, biological process, and pathway associations.
  - If you are uploading a list with values, a third column should be included with the values

Microsoft Excel file format is not accepted by the tool. To convert an Excel format (.xls or .xlsx) to a simple text format (.txt) - Open the file in Excel, under *File* menu, choose *Save As*, and under the *Format* option, choose *Tab Delimited Text*, and then click the *Save* button.

2.1.3 Step 2. Select organism

Because some identifiers, such as gene symbols, are not organism specific, it is important to select the organism that your IDs are from. However, you can select multiple organisms, especially when you run the functional classification, and select the organism on the result page. The organisms are organized as below. The first 12 organisms on the list are model organisms, followed by the rest organisms ordered alphabetically.

2.1.4 Step 3.Select Analysis

You can choose one of the following tools to analyze your list.

2.1.4.1 Functional classification viewed in gene list

This is the same tool as *Batch ID upload* in previous PANTHER releases.

Upload a list of gene/protein IDs, and obtain functional classifications with terms from Gene Ontology, PANTHER Protein Class, and PANTHER Pathway. The results are displayed on a gene list page (3.2.1).

2.1.4.2 Function classification viewed in pie chart

This tool is the same as above, but the results are displayed in a pie chart. See section 3.4.1 for details.

2.1.4.3 Statistical overrepresentation test

Use the binomial statistics tool to compare classifications of multiple clusters of lists to a reference list to statistically determine over- or under- representation of PANTHER classification categories. Each list is compared to the reference list using the binomial test (Cho and Campbell, TIGs 2000) [7] for each Gene Ontology, PANTHER protein class, or pathway term in PANTHER.

The default setting for this analysis is as below:

- Annotation data set: PANTHER GO-slim Biological Process
- Reference gene list: All genes in PANTHER database for the selected organism
- Correction: Bonferroni correction

You can reach the Result page (Section 2.1.4.3.1) by clicking the Submit button.

If you want to change the setting, you can uncheck the "Use default settings", and click the Submit button. You will reach the setting page (Fig. 2.3) that you can do the followings.

change Annotation Data Set, Reference Gene List, and add additional gene list(s) to analyze. You can analyze up to four lists with this tool. For example, the reference list may be the set of all genes in the experiment (ex: the list of all genes on the microarray), or the set of all genes in the genome being analyzed. The default is the set of all genes in the human genome. You may also upload your own list, or choose from one of the available lists such as the mouse and rat genomes. .

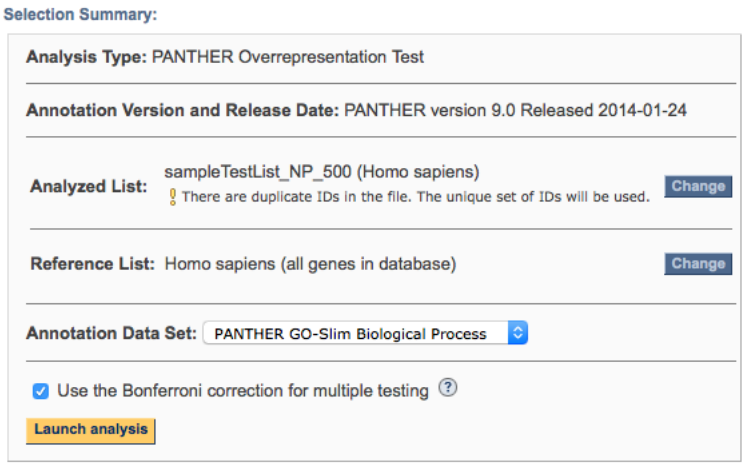

Selection Summary:
Analysis Type: PANTHER Overrepresentation Test
Annotation Version and Release Date: PANTHER version 9.0 Released 2014-01-24
Analyzed List: sampleTestList_NP_500 (Homo sapiens)
There are duplicate IDs in the file. The unique set of IDs will be used.
Change
Reference List: Homo sapiens (all genes in database)
Change
Annotation Data Set: PANTHER GO-Slim Biological Process
Use the Bonferroni correction for multiple testing
Launch analysis

Figure 2.3: *Modify settings for the overrepresentation test.*

- Change input gene list or upload additional lists - On the setting page (Fig. 2.3), you can click the "Change" button next to the Analyzed List to reach a page (Fig. 2.4) that allows you to do the followings:
  - The uploaded gene list appears on the right side of the page. You can delete a list by clicking the "x" icon in front of the list name (Figure 2.4).
  - You can also click the "Browse" button to upload additional gene lists. You can upload up to 4 gene lists for the analysis.

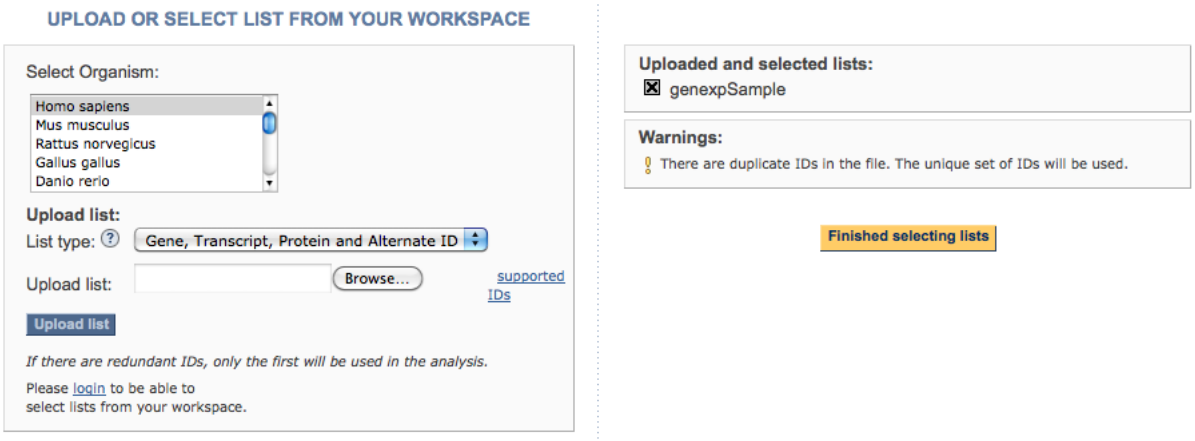

UPLOAD OR SELECT LIST FROM YOUR WORKSPACE
Select Organism:
Homo sapiens
Mus musculus
Rattus norvegicus
Gallus gallus
Danio rerio
Upload list:
List type:
Gene, Transcript, Protein and Alternate ID
Upload list:
Browse...
supported IDs
Upload list
If there are redundant IDs, only the first will be used in the analysis.
Please login to be able to select lists from your workspace.
Uploaded and selected lists:
genexpSample
Warnings:
There are duplicate IDs in the file. The unique set of IDs will be used.
Finished selecting lists

Figure 2.4: *Change input gene list or upload additional gene lists for the overrepresentation test.*

- Modify reference gene list and proceed - The reference list may be the set of all genes in the experiment (ex: the list of all genes on the microarray), or the set of all genes in the genome being analyzed. The default is the set of all genes in the human genome. You may also upload your own list, or choose from one of the available lists such as the mouse and rat genomes. To do so, from the setting page (Fig. 2.3), click the "Change" button next to the Reference List to reach the page (Fig. 2.4) as before that allows you to upload

a new reference list.

- Modify Annotation Data Set for the analysis - The default is the PANTHER GO-Slim Biological Process. You can change it to any data set described in Section 2.1.1 by using the drop-down menu next to the Annotation Data Set.

2.1.4.3.1 Analysis results

Analysis Summary: Please report in publication

Analysis Type: PANTHER Overrepresentation Test

Annotation Version and Release Date: PANTHER version 9.0 Released 2014-01-24

Analyzed List: sampleTestList_NP_500 (Homo sapiens) Change
There are duplicate IDs in the file. The unique set of IDs will be used.

Reference List: Homo sapiens (all genes in database) Change

Annotation Data Set: PANTHER GO-Slim Biological Process

Use the Bonferroni correction for multiple testing

Results

	Reference list	sampleTestList_NP_500
Mapped IDs:	21804	508
Unmapped IDs:	0	31

Export results View: -- Please select a chart to display --

Displaying only results with P<0.05; click here to display all results

	Homo sapiens (REF)	sampleTestList_NP_500				
PANTHER GO-Slim Biological Process	#	#	expected	Fold Enrichment	+/-	▲ P value
Unclassified	9422	87	219.52	.40	-	0.00E00
fatty acid metabolic process	188	45	4.38	> 5	+	3.45E-28
lipid metabolic process	902	86	21.02	4.09	+	5.22E-26
neuron-neuron synaptic transmission	90	30	2.10	> 5	+	2.17E-22
coenzyme metabolic process	101	31	2.35	> 5	+	3.91E-22
neurological system process	1212	93	28.24	3.29	+	1.28E-21
system process	1481	103	34.51	2.99	+	4.71E-21
regulation of vasoconstriction	51	20	1.19	> 5	+	6.27E-16
blood circulation	149	29	3.47	> 5	+	2.09E-15
multicellular organismal process	1798	103	41.89	2.46	+	4.22E-15
single-multicellular organism process	1798	103	41.89	2.46	+	4.22E-15
lipid transport	301	37	7.01	> 5	+	1.37E-13
fertilization	82	21	1.91	> 5	+	3.77E-13
acyl-CoA metabolic process	30	14	.70	> 5	+	7.08E-12
cell-cell signaling	835	58	19.45	2.98	+	5.87E-11
metabolic process	8613	281	200.67	1.40	+	9.39E-11
synaptic transmission	415	38	9.67	3.93	+	4.20E-10
cellular amino acid biosynthetic process	125	21	2.91	> 5	+	1.06E-09
anion transport	141	22	3.29	> 5	+	1.35E-09
localization	2636	116	61.41	1.89	+	2.34E-09
immune system process	1733	87	40.38	2.15	+	2.64E-09
transport	2564	113	59.74	1.89	+	4.30E-09
fatty acid beta-oxidation	35	12	.82	> 5	+	1.60E-08
primary metabolic process	7177	235	167.21	1.41	+	6.22E-08
response to toxic substance	61	14	1.42	> 5	+	7.61E-08
cell death	697	46	16.24	2.83	+	1.01E-07

Figure 2.5: *Results of the binomial distribution analysis.*

On the top of the results page (Fig. 2.5), there is a box summarizes the setting of the analysis. If the results are to be published, the information should be included in it. You can also make modifications of the setting in the box and rerun the analysis.

The analysis results are shown in a table with seven essential columns of data. By default, only the categories with p-value better than 0.05 are displayed. All results can be displayed by clicking the "click here to display all results" link.

1. The first column contains the name of the annotation data category, such as PANTHER GO-slim, GO or PANTHER pathway. If you are doing this analysis in terms of pathways, you can click on the pathway name to view the corresponding pathway diagram.

2. The second column contains the number of genes in the reference list that map to this particular annotation data category.

3. The third column contains the number of genes in your uploaded list that map to this annotation data category.

4. The fourth column contains the expected value (see Tips in blue below), which is the number of genes you would expect in your list for this category, based on the reference list.

5. The fifth column shows the Fold Enrichment of the genes observed in the uploaded list over the expected (number in your list divided by the expected number). If it is greater than 1, it indicates that the category is overrepresented in your experiment. Conversely, the category is underrepresented if it is less than 1.

6. The sixth column has either a + or -. A plus sign indicates over-representation of this category in your experiment: you observed more genes than expected based on the reference list (for this category, the number of genes in your list is greater than the expected value). Conversely, a negative sign indicates under-representation.

7. The seventh column is the p-values (see Tips below in blue) as determined by the binomial statistic. This is the probability that the number of genes you observed in this category occurred by chance (randomly), as determined by your reference list. A small p-value indicates that the number you observed is significant and potentially interesting. A cutoff of 0.05 is recommended as a starting point.

If you upload more than one list, the 3-6 columns are repeated for each list. The result is sorted based on p-values of the first list, but you can click on any column header to sort based on that column. The yellow triangle in front of the header indicates which column is sorted, and the orientation of the triangle indicates whether the sort is ascending or descending.

Tip: Binomial distribution test

Expected Value

The expected value is the number of genes you would expect in your list for a particular PANTHER category, based on the reference list. As an example, say we do do the analysis in terms of biological processes, and we are interested in determining if there is over- or under-representation of chromatin packaging and remodeling genes in your uploaded list.

As an example, there are 23,481 genes in your reference list (ex: the human genome). 125 of these genes map to the 'Chromatin packaging and remodeling' category. Based on this, 0.532% (125 divided by 23481) of the genes in the reference list are involved in chromatin packaging and remodeling.

Now your upload list contains 978 genes. Based on the reference list, we would expect that 5.21 genes (978 * 0.532%) in your uploaded list would be involved in chromatin packaging and remodeling.

If for this biological process you observe more genes in your uploaded list than expected, you have an over-representation (+) of genes involved in chromatin packaging and remodeling. If you observe fewer genes than expected, you have an under-representation (-).

P-Value calculated by the Binomial statistic

In the example above we expected 5.21 genes in your uploaded list to be involved in chromatin packaging and remodeling. Say that we actually observe 7 genes involved in this biological process. This is very similar to 5.21, so you would expect a large p-value (the p-value would be 0.27). Alternatively, if 15 of your uploaded genes map to this process, this is very different than the expected p-value, so you would expect a small, significant p-value (the p-value would be 3.3E-04). This small p-value indicates that the result is non-random and potentially interesting, and worth looking at in closer detail. A p-value cutoff of 0.05 is recommended as a start point. The binomial statistic is a commonly used statistic (it can be found in any statistics book). In the binomial test we assume that under the NULL hypothesis, genes in the uploaded list are sampled from the same general population as genes from the reference set, i.e. the probability p(C) of observing a gene from a particular category C in the uploaded list is the same as in the reference list. We first estimate the probability p(C) from the reference set assuming that it is large and representative:

p(C)=n(C)/N,

where n(C) is the number of genes mapped to category C, and N is the total number of genes in the reference set.

We then use the above estimate to find the p-value: the probability of observing k(C) genes (or a more extreme number) in the uploaded list of size K. Under the NULL hypothesis, the number of genes of mapped to C is distributed binomially with probability parameter p(C) and thus the p-value would be

$$p-value=\sum\binom{K}{k}p(c)^{k}(1-p(c))^{K-k}$$

where the sum runs from k(C) to K in the case of over-representation (i.e. when the number of observed genes k(C) is greater than expected p(C)*K under the NULL hypothesis), and 0 to k(C), in the case of under-representation (i.e. when k(C) is smaller than p(C)*K).

When developing this analysis tool, we tested using both the Chi-Square and Binomial statistical tests. We decided to use the Binomial, since the Chi-Square is not as accurate when the population sizes or the expect number is small.

2.1.4.3.2 *View pathway diagram*

If you are analyzing pathways, you can click the pathway name in the first column and view pathway diagram with your IDs overlay on the diagram. Each list has an assigned color as shown on the top of the result page (Figure 2.5). The drop-down menu allows you to select a color of choice for each list.

Figure 2.6 shows an example of such a diagram. Red and green indicate the genes appearing in each of your list. Yellow indicates that the pathway component is in more than one of the uploaded list (not including the reference list). Gray indicates the components not in your list, but only in the reference list.

The diagram can be manipulated as described in section 3.4.2. The colored image can only be exported as a png file using the "File-¿Export image" function in the applet menu.

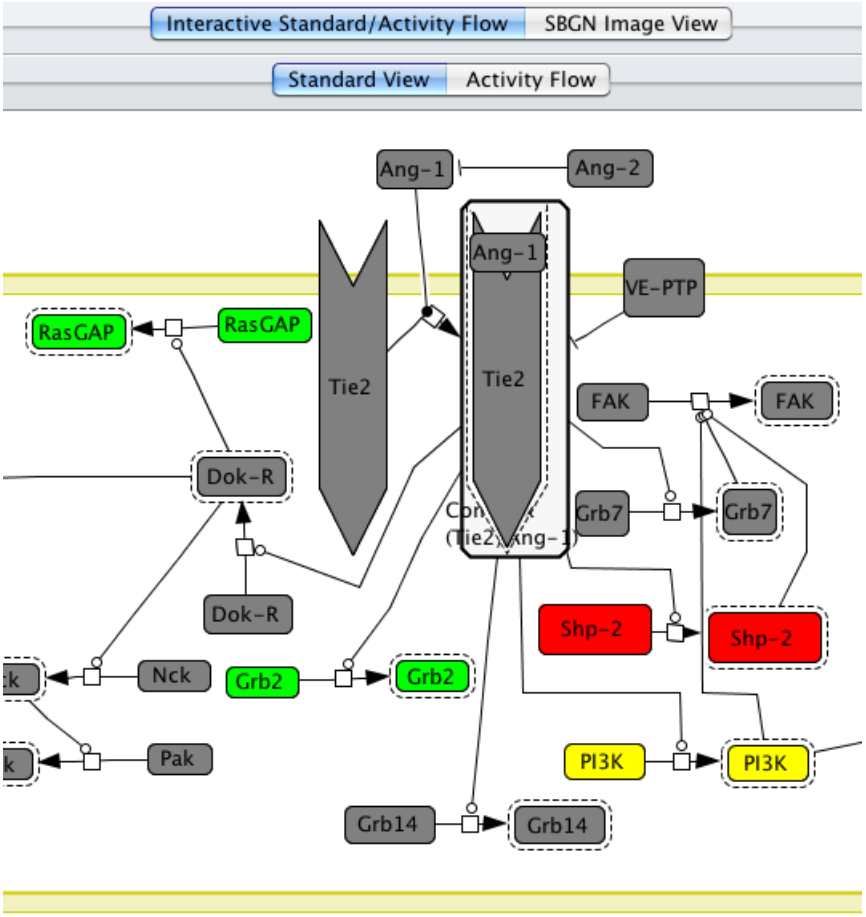


Figure 2.6: *View results in pathways. Each list id colored by a different color. A total of 4 lists can be uploaded.*

2.1.4.4 Statistical enrichment test

For each molecular function, biological process, cellular component, PANTHER protein class, or pathway term in PANTHER, the genes associated with that term are evaluated according to the likelihood that their numerical values were drawn randomly from the overall distribution of values. The Mann-Whitney U Test (Wilcoxon Rank-Sum Test) is used to determine the P-value that, say, the chromatin packaging and remodeling genes have random values relative to overall list of values that were input.

To use this test, make sure that the numerical values are included in the input file. See File format for more details.

This approach has been used by our group (Clark et al., Science 302: 1960, 2003) [8] and is similar to a method from Eric Lander's group (Mootha et al., Nature Genetics 34: 267, 2003) [9], to find weakly coordinated shifts that elude methods based on defining strict cutoffs in the data, e.g. only focusing on genes whose expression has changed by over 1.5- or 2-fold.

For the rank-sum test, it is important to provide values for as many genes as possible (subject to noise level and reliability) so that randomness can be properly assessed across the experiment.

The default setting for this analysis is as below:

- Annotation data set: PANTHER GO-slim Biological Process
- Correction: Bonferroni correction

You can reach the Result page by clicking the Submit button.

If you want to change the setting, you can also uncheck the "Use default setting", and click the Submit button. You will reach the setting page (Fig. 2.7) that you can change Annotation Data Set.

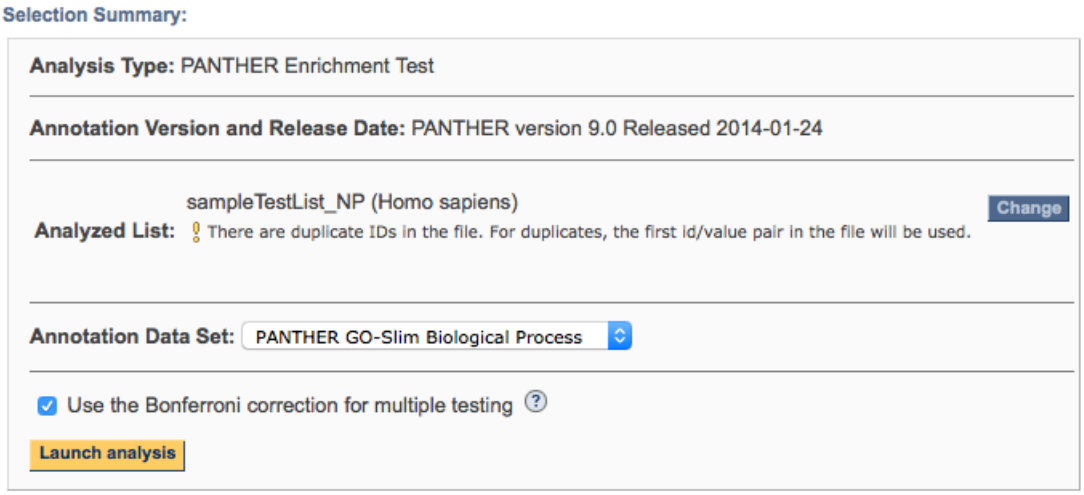
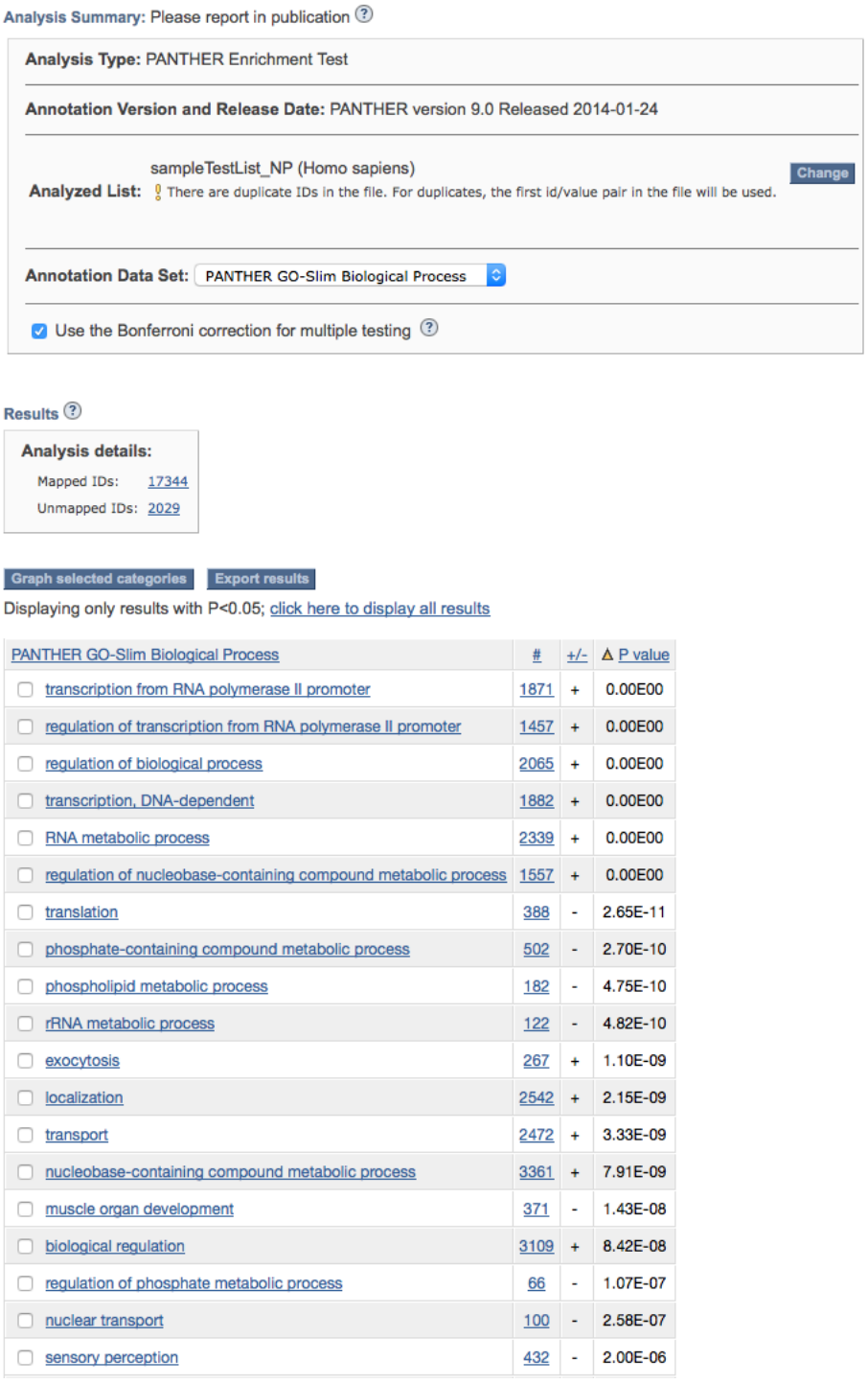

Selection Summary:
Analysis Type: PANTHER Enrichment Test
Annotation Version and Release Date: PANTHER version 9.0 Released 2014-01-24
sampleTestList_NP (Homo sapiens)
Change
Analyzed List:
There are duplicate IDs in the file. For duplicates, the first id/value pair in the file will be used.
Annotation Data Set:
PANTHER GO-Slim Biological Process
Use the Bonferroni correction for multiple testing
Launch analysis

Figure 2.7: *Modify settings for the enrichment test.*

2.1.4.4.1 Results



Analysis Summary: Please report in publication

Analysis Type: PANTHER Enrichment Test

Annotation Version and Release Date: PANTHER version 9.0 Released 2014-01-24

Analyzed List: sampleTestList_NP (Homo sapiens) Change

There are duplicate IDs in the file. For duplicates, the first id/value pair in the file will be used.

Annotation Data Set: PANTHER GO-Slim Biological Process

Use the Bonferroni correction for multiple testing

Results

Analysis details:

Mapped IDs: 17344

Unmapped IDs: 2029

Graph selected categories | Export results

Displaying only results with P<0.05; click here to display all results

PANTHER GO-Slim Biological Process	#	+/-	Δ P value
transcription from RNA polymerase II promoter	1871	+	0.00E00
regulation of transcription from RNA polymerase II promoter	1457	+	0.00E00
regulation of biological process	2065	+	0.00E00
transcription, DNA-dependent	1882	+	0.00E00
RNA metabolic process	2339	+	0.00E00
regulation of nucleobase-containing compound metabolic process	1557	+	0.00E00
translation	388	-	2.65E-11
phosphate-containing compound metabolic process	502	-	2.70E-10
phospholipid metabolic process	182	-	4.75E-10
rRNA metabolic process	122	-	4.82E-10
exocytosis	267	+	1.10E-09
localization	2542	+	2.15E-09
transport	2472	+	3.33E-09
nucleobase-containing compound metabolic process	3361	+	7.91E-09
muscle organ development	371	-	1.43E-08
biological regulation	3109	+	8.42E-08
regulation of phosphate metabolic process	66	-	1.07E-07
nuclear transport	100	-	2.58E-07
sensory perception	432	-	2.00E-06

Figure 2.8: *The Mann-Whitney test result page..*

On the top of the results page (Fig. 2.8), there is a box summarizes the setting of the analysis. If the results are to be published, the information should be included in it. You can also make modifications of the setting in the box and rerun the analysis.

On the results page, there is a table with four essential columns of data. By default, only the categories with p-value better than 0.05 are displayed. All results can be displayed by clicking

the "click here to display all results" link.

1. The first column contains the name of the annotation data category, such as PANTHER GO-slim, GO or PANTHER pathway. If you are doing this analysis in terms of pathways, you can click on the pathway name to view the pathway diagram. The genes in the pathway diagram are colored according to the gene expression value, and the rules for this can be specified by clicking on the 'Specify color ranges' button.

2. The second column contains the number of genes that map to this particular classification category.

3. The third column has either a + or -. A plus sign indicates that for this category, the distribution of values for your uploaded list is shifted towards greater values than the overall distribution of all genes that were uploaded. A negative sign indicates that the uploaded list is shifted towards smaller values than the overall list.

4. The fourth column contains the p-value as calculated from the (Mann-Whitney U Test (Wilcoxon Rank-Sum test). A large p-value indicates that the genes for this category have a distribution that is similar to randomly choosing genes from the overall distribution. In other words, the values of the uploaded genes for this category have a similar distribution to the overall list of values that were input. A small, significant p-value indicates that the distribution for this category is non-random and different than the overall distribution. A cutoff of 0.05 is recommended as a starting point.

To have a visual representation of these distributions, select the checkboxes of the categories of interest, and click on the "Graph selected categories" button near the top of the page. Figure 2.9 shows an example of the graph. The x-axis is your uploaded value. The y-axis is the cumulative fraction. In other words, if you look at the data point x=1.4, y=.75 this means that 75% of your uploaded values have a value of 1.4 or smaller.

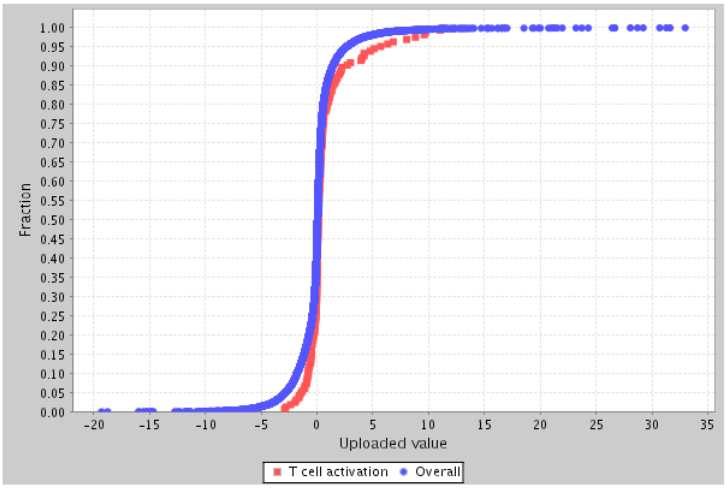


Figure 2.9: *Mann-Whitney test results shown in graph view.*

The Mann-Whitney U Test (Wilcoxon Rank-Sum test) Statistic

To perform the rank sum test, first the values of the genes that map to a given category are combined with the overall list of values that were input. Then, all the values are ranked from smallest to largest, with the smallest value getting a rank of 1. If multiple values are identical, the average of the ranks for these values is used.

Then the rank sum is calculated for this category, by summing up the ranks for all of the genes that map to this category. The average rank, R1 is then calculated by dividing the rank sum by the number of genes, n1, that map to the category. Likewise, the rank sum is calculated for the list of all IDs uploaded, and the average rank, R2, is calculated by dividing the rank sum by the total number of genes uploaded, n2.

Next, the Mann Whitney U statistic is calculated for both populations:

U1 = n1* n2 + (n1 * (n1 + 1)) / 2 - R1
U2 = n2* n2 + (n1 * (n2 + 1)) / 2 - R2

The larger of these two values is the Mann Whitney U-statistic, U, whose distribution for small sample sizes can be found in most statistic books. In our case, our application is for large sample sizes, so we use the normal approximation:

Z-score = (U- (n1* n2)/2)/sqrt(n1*n2*(n1+n2+1)/12).

It follows that the p-value is the integral under the standard normal density.

2.1.4.4.2 *View results in pathway diagram*

If you select to analyze the data in PANTHER pathway, you can click the pathway name from the column 1 of the result page, and view the results in pathway diagram in "head map" (Figure 2.10).

The color range reflects the expression value of your genes mapped to the pathway component. You can define the color range by selecting "Specify color ranges for pathway diagram" button on the result page (Figure 2.8). First there are two color ranges to select. Select equal distributions (default), and the ranges will be determined by sorting your gene values, and then splitting the values into six equal distributions. Select fold changes, and standard gene expression fold change ranges will be used.

Some pathway components have multiple genes mapped to them. You can see color ranges for these components that are determined by

- median of the gene values *(default)*
- average of the gene values
- maximum gene value
- minimum gene value

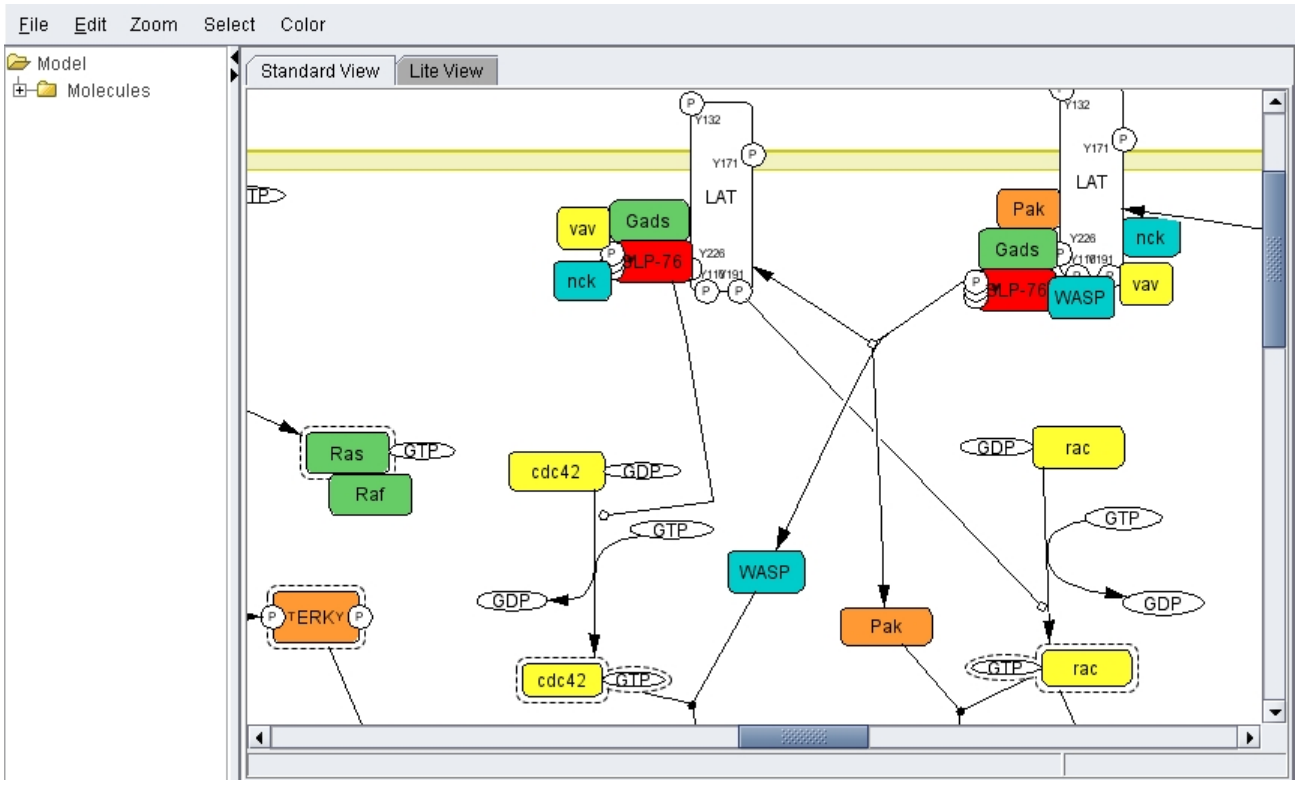


Figure 2.10: *Mann-Whitney test results shown in pathway diagram in a "head map".*

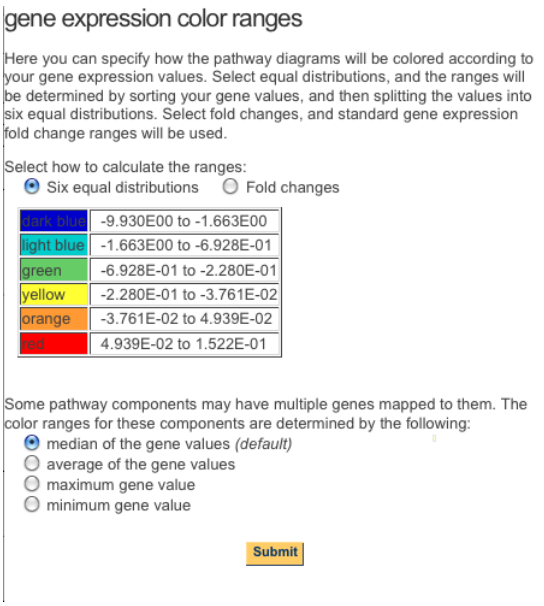

gene expression color ranges
Here you can specify how the pathway diagrams will be colored according to your gene expression values. Select equal distributions, and the ranges will be determined by sorting your gene values, and then splitting the values into six equal distributions. Select fold changes, and standard gene expression fold change ranges will be used.
Select how to calculate the ranges:
Six equal distributions
Fold changes
dark blue -9.930E00 to -1.663E00
light blue -1.663E00 to -6.928E-01
green -6.928E-01 to -2.280E-01
yellow -2.280E-01 to -3.761E-02
orange -3.761E-02 to 4.939E-02
red 4.939E-02 to 1.522E-01
Some pathway components may have multiple genes mapped to them. The color ranges for these components are determined by the following:
median of the gene values (default)
average of the gene values
maximum gene value
minimum gene value
Submit

Figure 2.11: *Color code for the head map.*

2.2 Browse

This is the second menu tab on the PANTHER home page.

The PANTHER Ontology Browser (or PANTHER Prowler) is a highly interactive Java

applet used to browse the PANTHER ontologies, make selection(s) , and retrieve results for different data associated with the ontology terms, such as individual genes or families and subfamilies of proteins.

The Ontology Browser is composed of 4 panels (Figure 2.12).

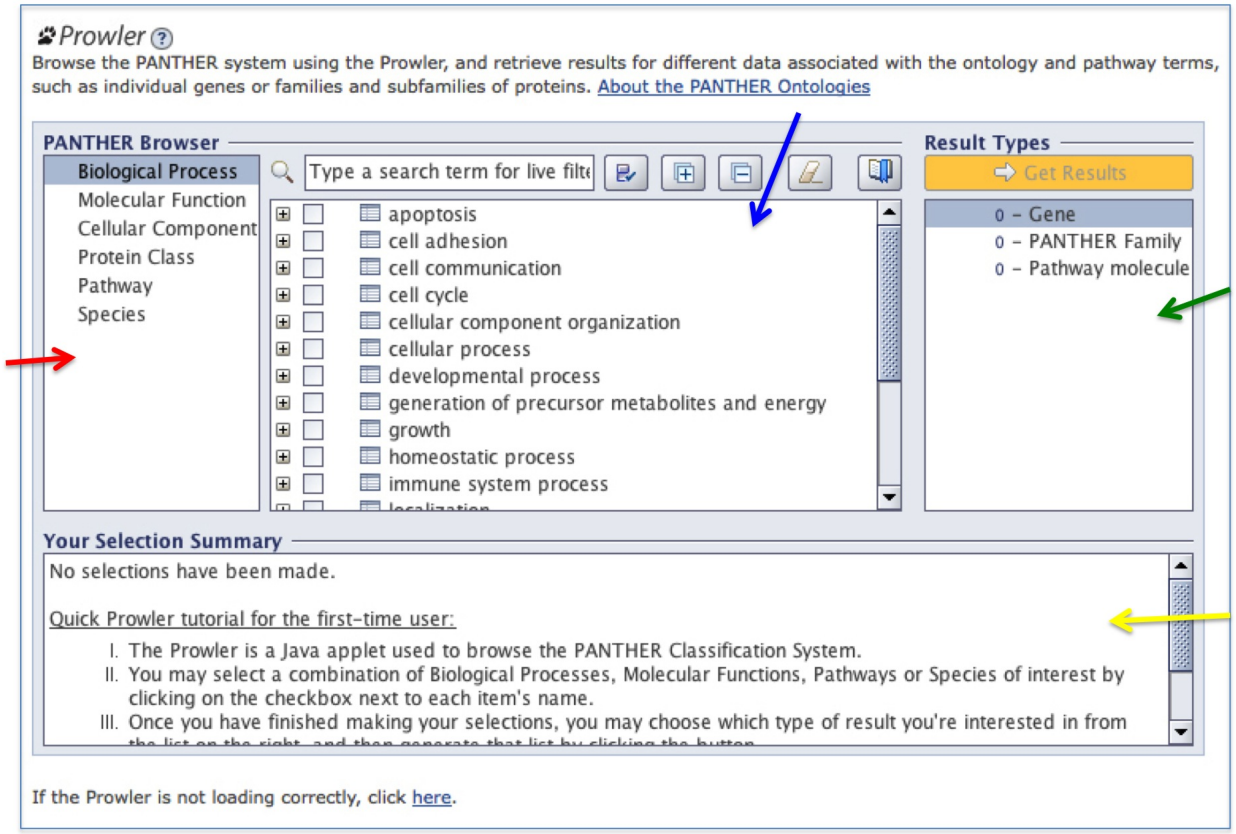

Prowler
Browse the PANTHER system using the Prowler, and retrieve results for different data associated with the ontology and pathway terms, such as individual genes or families and subfamilies of proteins. About the PANTHER Ontologies
PANTHER Browser
Biological Process
Molecular Function
Cellular Component
Protein Class
Pathway
Species
Type a search term for live filte
apoptosis
cell adhesion
cell communication
cell cycle
cellular component organization
cellular process
developmental process
generation of precursor metabolites and energy
growth
homeostatic process
immune system process
Result Types
Get Results
0 - Gene
0 - PANTHER Family
0 - Pathway molecule
Your Selection Summary
No selections have been made.
Quick Prowler tutorial for the first-time user:
I. The Prowler is a Java applet used to browse the PANTHER Classification System.
II. You may select a combination of Biological Processes, Molecular Functions, Pathways or Species of interest by clicking on the checkbox next to each item's name.
III. Once you have finished making your selections, you may choose which type of result you're interested in from
If the Prowler is not loading correctly, click here.

Figure 2.12: *The PANTHER Prowler*

- Ontology Type (red arrow) — On the left side of the Browser is the Ontology Type panel. There are **five** ontology types you can choose.
  - Biological Process – this is a GO Slim *biological process* ontology
  - Molecular Function – this is a GO Slim *molecular function* ontology
  - Cellular Component – this is a GO Slim *cellular component* ontology
  - Protein Class – this is PANTHER ontology to categorize protein families
  - Pathway

  The last choice is *species*. You can choose the species when querying over certain result sets such as genes, to narrow your search. There are 48 total species that you can choose, with 12 model organism species listed first, and the 36 remaining in alphabetical order. The default is all species.

- Ontology (blue arrow) — This panel is in the middle of the prowler, and it changes based on the selection of the ontology type. You can click on the + to expand the next level of

the ontology terms. The details of each ontology term can be viewed by mouse-over the name, or by clicking the little icon right in front of the name.

On the top of the panel, there are a few special utilities:

- Search. You can type the term in the *Type the search term for live filtering* box, and only the ontology terms containing the search will be displayed.
- Expand to selections. Clicking this button will expand only the paths to the selected ontology terms.
- Expand all. Clicking this button will expand the ontology so that all child terms (subcategories) are shown.
- Collapse all. Clicking this button will collapse the ontology so that only the topmost (general) terms are shown.
- Clear all selections. Clicking this button will clear all selections of all types of ontologies.
- Bookmark. Clicking this button will cause the page to reload with a special address (URL). This address will have all your selection encoded. If you save this address to your bookmark or favorites, you can return to your current selection at any time. This address is also suitable for cut and paste into your documents, email messages, or reference documentation.

- Results Type (green arrow) — This panel is on the right side, and it displays the number of genes, PANTHER families, and pathway molecule classes (components) that matches your selections. You can choose one of the above result types, and click the *Get Results* button. The results will be displayed as a list page (Section 3.2).

- Selection Summary (yellow arrow) — This panel is in the bottom of the Prowler. Since you are allowed to make selections from multiple ontologies, a summary of the query results from your selections is displayed in this panel.

Browse the PANTHER ontologies
Figure 2.13 shows an example of making selections using the PANTHER Prowler.

1. Select an ontology on the *Ontology type* panel.

2. Select the ontology you are interested in (left panel). Typing in the text field will filter the list of possible selections. Click on a + (red arrow) to expand a category to view subcategories, and select categories by clicking on the checkbox next to the name. Brief details about each category can be seen by mousing-over a name, or full details by clicking on the icon to the left of a name. For pathways, full details include a detailed diagram of the molecular interactions and reactions in the pathway.

3. More than one ontology can be selected. When multiple selections are made, it means that results will meet the criteria from all the selections. In this example, selections are made in both *Molecular Function* and *Pathway*, and the selected ontology can be seen as marked in the Ontology panel (red oval in Figure 2.13).

4. Review your selections in the Selection Summary panel (green arrow). For example, in this example, one selects *Human* from the species ontology as well as *Protein kinase activity* from the molecular function ontology and *Apoptosis signaling pathway* in pathway to retrieve all genes that have protein kinase activity that involve in apoptosis pathway in the human genome.

5. Human curators have associated ontology terms with genes, PANTHER families, subfamilies and pathways To see these associations, select the result type in the right panel and press the orange *Get Results* button (blue arrow). It will take you to the list pages.

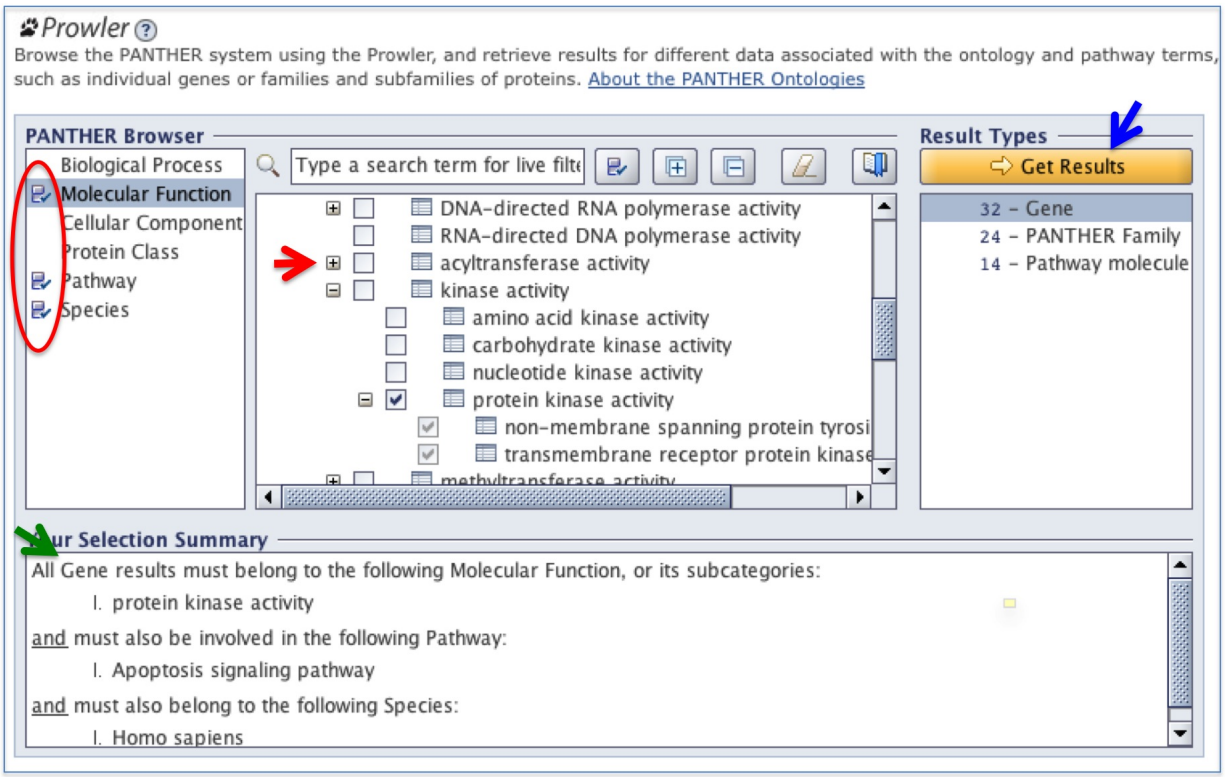

Prowler
Browse the PANTHER system using the Prowler, and retrieve results for different data associated with the ontology and pathway terms, such as individual genes or families and subfamilies of proteins. About the PANTHER Ontologies
PANTHER Browser
Biological Process
Molecular Function
Cellular Component
Protein Class
Pathway
Species
Type a search term for live filte
DNA-directed RNA polymerase activity
RNA-directed DNA polymerase activity
acyltransferase activity
kinase activity
amino acid kinase activity
carbohydrate kinase activity
nucleotide kinase activity
protein kinase activity
non-membrane spanning protein tyrosi
transmembrane receptor protein kinase
Result Types
Get Results
32 - Gene
24 - PANTHER Family
14 - Pathway molecule
ur Selection Summary
All Gene results must belong to the following Molecular Function, or its subcategories:
l. protein kinase activity
and must also be involved in the following Pathway:
l. Apoptosis signaling pathway
and must also belong to the following Species:
l. Homo sapiens

Figure 2.13: *An example showing selections using Prowler*

2.3 Sequence Search

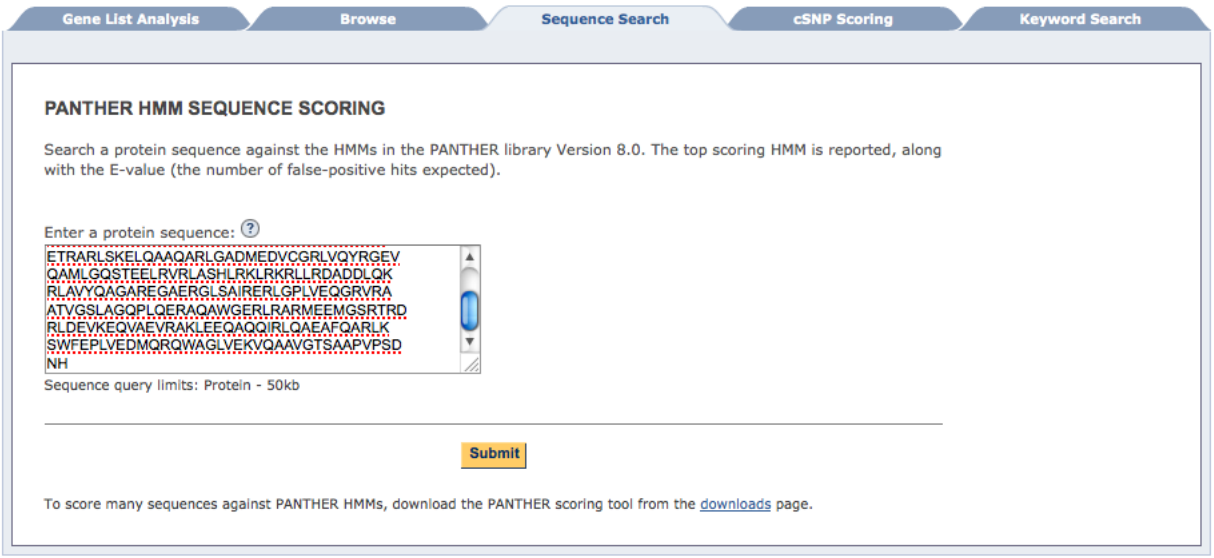

Gene List Analysis
Browse
Sequence Search
cSNP Scoring
Keyword Search
PANTHER HMM SEQUENCE SCORING
Search a protein sequence against the HMMs in the PANTHER library Version 8.0. The top scoring HMM is reported, along with the E-value (the number of false-positive hits expected).
Enter a protein sequence:
ETRARLSKELQAAQARLGADMEDVCGRLVQYRGEV
QAMLGQSTEELRVRLASHLRKLRKRLLRDADDLQK
RLAVYQAGAREGAERGLSAIRERLGPLVEQGRVRA
ATVGSLAGQPLQERAQAWGERLRARMEEMGSRTRD
RLDEVKEQVAEVRAKLEEQAQQIRLQAEAFQARLK
SWFEPLVEDMQRQWAGLVEKVQAAVGTSAAPVPSD
NH
Sequence query limits: Protein - 50kb
Submit
To score many sequences against PANTHER HMMs, download the PANTHER scoring tool from the downloads page.

Figure 2.14: *Sequence Search - Search protein sequence against the PANTHER HMM library.*

The function can be accessed from the 3rd tab from the PANTHER home page (Figure 2.14), or the "interactive score of new sequences against PANTHER HMMs" link on the *Family and HMMs* subject home page (section 3.1.2). This online functionality only allows you to submit one sequence at a time. For multiple sequence scoring, you need to download the PANTHER scoring tool from the PANTHER ftp site at `http://www.pantherdb.org/downloads/`.

To score, just enter the amino acid sequence into the box, and click the "Submit" button. Here is how the result page is read (Figure 2.15).

- PANTHER hit: Only the top hit HMM is reported here. The family or subfamily id and name of the top hit are displayed on the page. Clicking the name will take you to the family/subfamily detail page (section 3.3.2).
- PANTHER score: The score of the PANTHER hit. The green dots next to the score indicates how closely related the protein is to the model. There are three categories:
  - closely related (indicated by 3 greet dots): if the score is better than E-23 (very likely to be a correct functional assignment)
  - related (indicated by 2 green dots): if the score is better than E-11, but worse than E-23 (molecular function likely to be the correct but biological process/pathway less certain)
  - distantly related (indicated by 1 green dot): if the score is better than E-3, but worse than E-11 (protein is evolutionarily related but function may have diverged)

PANTHER HMM SEQUENCE SCORING RESULTS

The top scoring HMM is reported, along with the E-value (the number of expected false-positive hits expected). If the E-value is less than 1e-3, no hits are reported.

PANTHER Hit: APOLIPOPROTEIN E (PTHR18976:SF2)
HMM E-value score: 2.2e-155

```
Sequence Domain  seq-f seq-t    hmm-f hmm-t      score  E-value
-------- ------- ----- -----    ----- -----      -----  -------
sequence   1/1       1   317 []     1   331 [.   527.0 2.2e-155

Alignments of top-scoring domains:
sequence: domain 1 of 1, from 1 to 317: score 527.0, E = 2.2e-155
                   *->lmkllwalmkvlllallvslLagcqakmkglavllalalltgcqars
                              mkvl++allv++Lagcqak++++++++++++l+ +q++
    sequence     1    --------MKVLWAALLVTFLAGCQAKVEQAVETEPEPELR-QQTE- 37

                   lpladepkqrWEealdrFwdYlrelqtladdvqeelkssQiskELdtLie
                       +++ qrWE+al+rFwdYlr++qtl+++vqeel+ssQ+++EL++L++
    sequence    38 ----WQSGQRWELALGRFWDYLRWVQTLSEQVQEELLSSQVTQELRALMD 83

                   DTMtELkaYkeeLeeqLtPvaeetrarlskelqalqarLgaDMeDakeRl
                   +TM+ELkaYk+eLeeqLtPvaeetrarlskelqa+qarLgaDMeD+  Rl
    sequence    84 ETMKELKAYKSELEEQLTPVAEETRARLSKELQAAQARLGADMEDVCGRL 133

                   tqYrgElqamleQnteevRaRvssylrKLrKRLlkDaeeLqkrlavYqag
                   +qYrgE+qaml+Q+tee+R R++s+lrKLrKRLl+Da++LqkrlavYqag
    sequence   134 VQYRGEVQAMLGQSTEELRVRLASHLRKLRKRLLRDADDLQKRLAVYQAG 183

                   aregaergvealRerlePlleqgqdraqaklgslaealksQakelkerle
                   aregaerg++a+Rerl+Pl+eqg+ r +a++gsla +
    sequence   184 AREGAERGLSAIRERLGPLVEQGRVR-AATVGSLAGQ------------- 219

                   sqaedlkerleataeelrgrLeeagerlrdrLdevreqieevkakvee..
                       +l+er++a++e+lr+r+ee+g+r+rdrLdev+eq++ev+ak+ee++
    sequence   220 ----PLQERAQAWGERLRARMEEMGSRTRDRLDEVKEQVAEVRAKLEEqa 265

                   .................FePLvedlqrQWagLvEKvQaavkpsaaetPas
                   ++ ++++++ ++++   FePLved+qrQWagLvEKvQaav +saa++P s
    sequence   266 qqirlqaeafqarlkswFEPLVEDMQRQWAGLVEKVQAAVGTSAAPVP-S 314

                   dnh<-*
                   dnh
    sequence   315 DNH    317
```

Figure 2.15: *The score result file from the PANTHER HMM score page.*

2.4 cSNP Scoring

This tool estimates the likelihood of a particular nonsynonymous (amino-acid changing) coding SNP to cause a functional impact on the protein. It calculates the subPSEC (substitution position-specific evolutionary conservation) score based on an alignment of evolutionarily related proteins, as described in Brunham et al., 2005 [10], Thomas et al., 2003 [11] and Thomas and Kejariwal, 2004 [12].

2.4.1 Input protein and substitution data

Paste the wildtype protein sequence in the first box (FASTA format, or just the protein sequence itself), and then enter the substitution(s) relative to this input sequence in the standard amino acid substitution format, e.g. A265V. Multiple substitutions should be separated by a tab, space, or return (Figure 2.16).

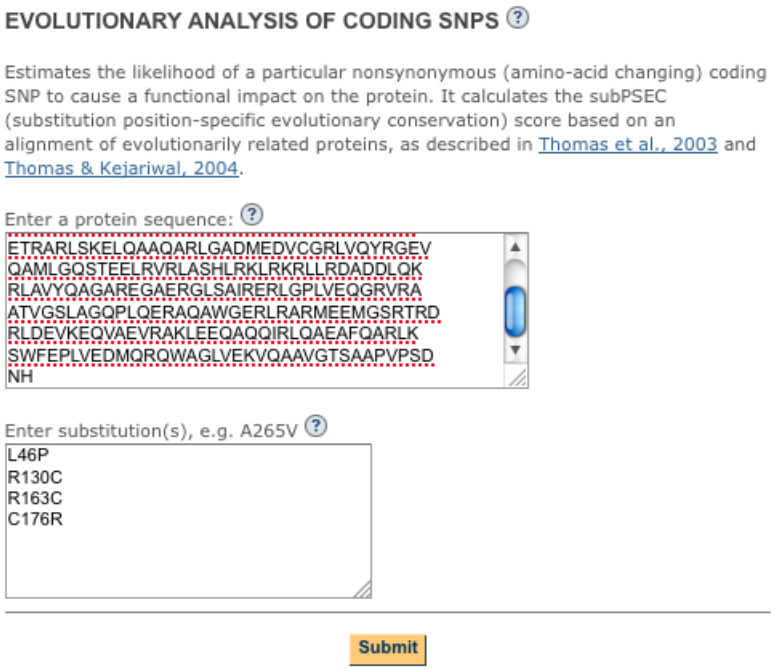

EVOLUTIONARY ANALYSIS OF CODING SNPS
Estimates the likelihood of a particular nonsynonymous (amino-acid changing) coding SNP to cause a functional impact on the protein. It calculates the subPSEC (substitution position-specific evolutionary conservation) score based on an alignment of evolutionarily related proteins, as described in Thomas et al., 2003 and Thomas & Kejariwal, 2004.
Enter a protein sequence:
ETRARLSKELQAAQARLGADMEDVCGRLVQYRGEV
QAMLGQSTEELRVRLASHLRKLRKRLLRDADDLQK
RLAVYQAGAREGAERGLSAIRERLGPLVEQGRVRA
ATVGSLAGQPLQERAQAWGERLRARMEEMGSRTRD
RLDEVKEQVAEVRAKLEEQAQQIRLQAEAFQARLK
SWFEPLVEDMQRQWAGLVEKVQAAVGTSAAPVPSD
NH
Enter substitution(s), e.g. A265V
L46P
R130C
R163C
C176R
Submit

Figure 2.16: *Enter sequence and SNP substitutions.*

Note that the sequence numbering must be correct relative to the sequence pasted in the box and not necessarily the numbering found in the literature. For example, APOE4 is listed as containing C112R in the literature, but this requires removing the signal peptide from the SWISS-PROT entry for APE_HUMAN, or entering the substitution as C130R.

2.4.2 Results of cSNP analysis tool

2.4.2.1 Protein score

The input protein sequence is scored against the HMMs in the PANTHER library. The alignment to the HMM with the most significant score is used for the analysis. Proteins that scored greater than 1e-23 are excluded from the analysis, as the alignments are less reliable.

2.4.2.2 subPSEC score and Pdeleterious

The subPSEC (substitution position-specific evolutionary conservation) score estimates the likelihood of a functional effect from a single amino acid substitution. It is the negative logarithm of the probability ratio of the wild-type and mutant amino acids at a particular position. PANTHER subPSEC scores are continuous values from 0 (neutral) to about -10 (most likely to be deleterious). -3 is the previously identified cutoff point for functional significance.

A cutoff of -3 corresponds to a 50% probability that a score is deleterious. From this, the probability that a given variant will cause a deleterious effect on protein function is estimated by Pdeleterious, such that a subPSEC score of -3 corresponds to a Pdeleterious of 0.5.

For more details and to see the corresponding equations, please see our most recent publication Brunham et al., 2005 [10]. Additional information is in Thomas et al., 2003 [11] and Thomas and Kejariwal, 2004 [12].

2.4.2.3 Multiple sequence alignment

Click on the link on the number of the multiple sequence alignment (MSA) position to view the column in the MSA where the substitution occurs. The column is highlighted in red, and the subfamilies (rows in the MSA) that were used to gather the statistics for amino acid probabilities

are highlighted in blue. In the MSA view, click on a blue diamond (a subfamily node) in the tree to hide or show sequences in the subfamily.

2.4.2.4 Number of independent counts (NIC)

NIC (number of independent counts) is an estimate of the number of independent observations used to calculate the amino acid probabilities. The probabilities are calculated from a combination of prior knowledge (e.g. that isoleucine often substitutes for valine) and observations, so the larger NIC, the more the probabilities rely on the amino acids observed in the multiple sequence alignment. position does not align to the HMM If the substitution occurs at a position that does not appear in the multiple sequence alignment; a subPSEC score cannot be generated and the output will return the text string "position does not align to the HMM", indicating that the substitution occurs at a position that is inserted relative to the consensus HMM for the given HMM. In most cases, these positions are not modeled by the HMMs simply because they do not appear in most of the related sequences; as a result, substitutions at inserted positions are not generally likely to be deleterious.

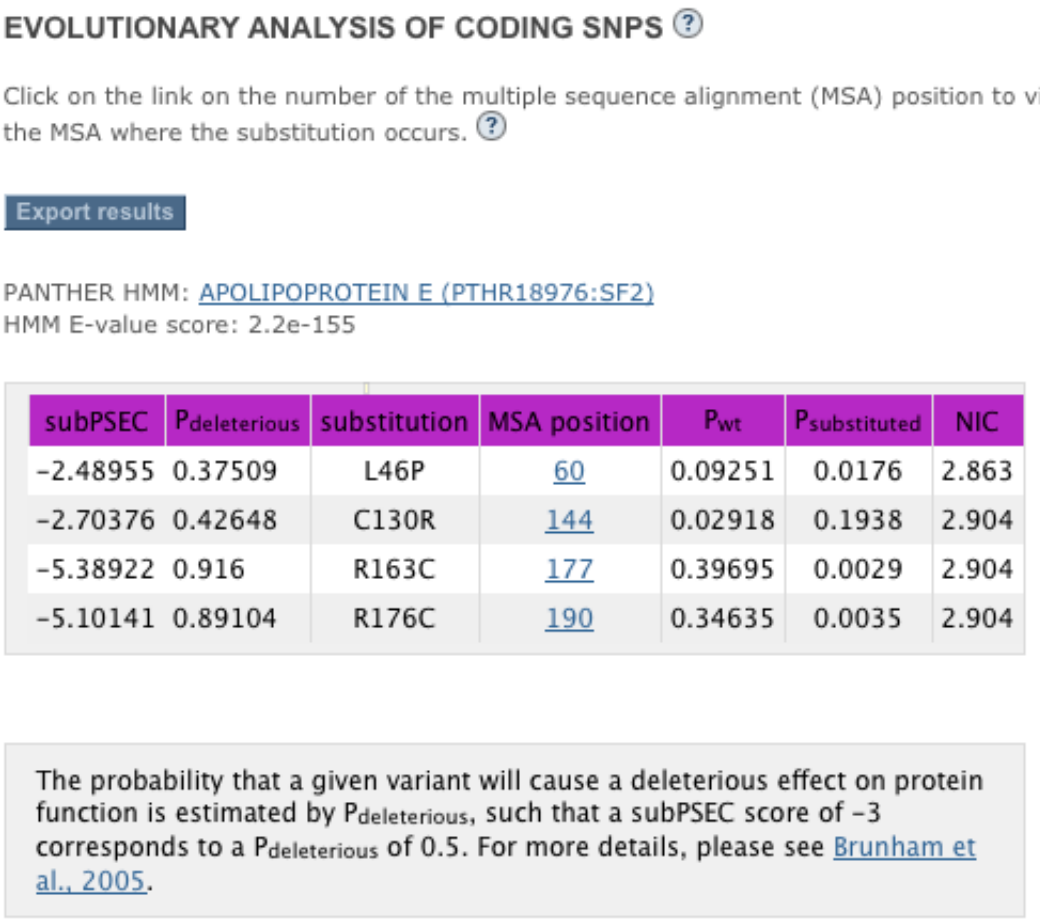

EVOLUTIONARY ANALYSIS OF CODING SNPS
Click on the link on the number of the multiple sequence alignment (MSA) position to vi
the MSA where the substitution occurs.
Export results
PANTHER HMM: APOLIPOPROTEIN E (PTHR18976:SF2)
HMM E-value score: 2.2e-155
subPSEC | Pdeleterious | substitution | MSA position | Pwt | Psubstituted | NIC
-2.48955 | 0.37509 | L46P | 60 | 0.09251 | 0.0176 | 2.863
-2.70376 | 0.42648 | C130R | 144 | 0.02918 | 0.1938 | 2.904
-5.38922 | 0.916 | R163C | 177 | 0.39695 | 0.0029 | 2.904
-5.10141 | 0.89104 | R176C | 190 | 0.34635 | 0.0035 | 2.904
The probability that a given variant will cause a deleterious effect on protein function is estimated by Pdeleterious, such that a subPSEC score of -3 corresponds to a Pdeleterious of 0.5. For more details, please see Brunham et al., 2005.

Figure 2.17: *cSNP analysis results page.*

2.5 Keyword Search

This is the last menu tab on the PANTHER home page.

To perform a keyword search, simply enter the search text in the "Enter Keyword(s)" box (Figure 2.18). Here are some tips about the search.

- The search term can be an identifier, a word or a phrase (multiple terms).
- The search looks for the exact match of the entire word. For example, if you enter the term "hexo", the search will not find "hexokinase". Therefore, if you are not sure about the word, a wild card (*) can be used. In this case, if you enter "hexo*", the search will find "hexokinase".
- The search looks for the search term(s) in all fields of the selected subject. Some of these fields may not be visible to you on the result list page, such as synonyms, definitions.

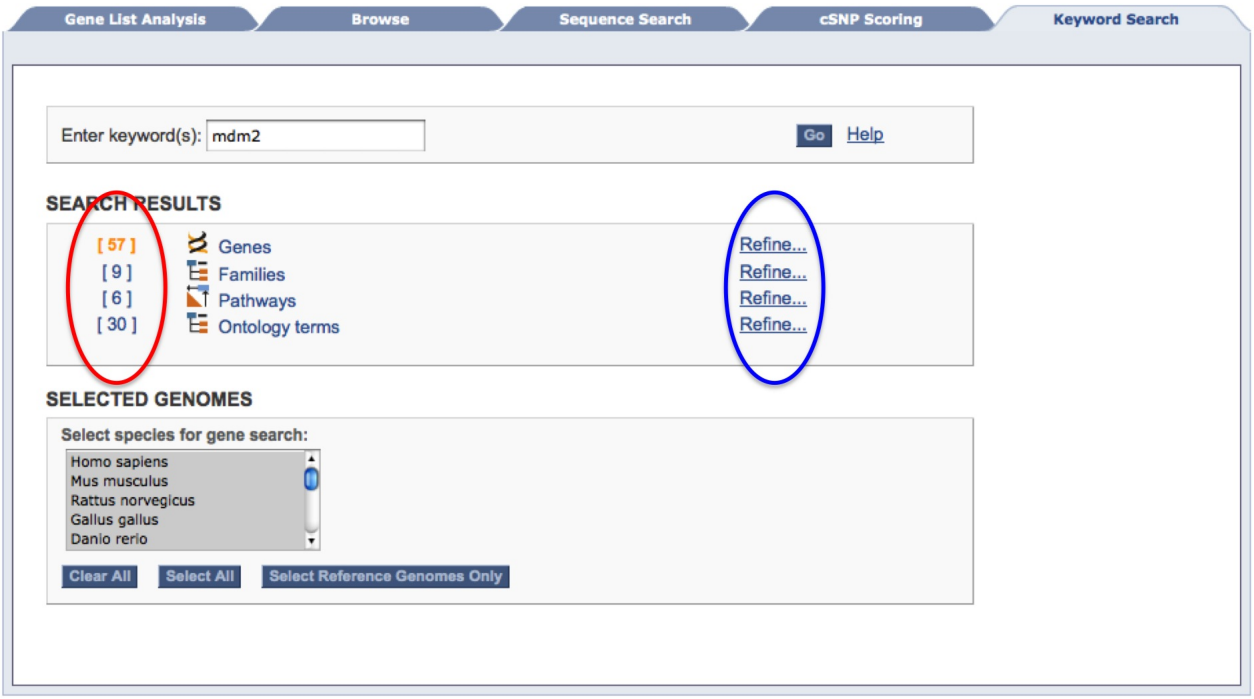


Figure 2.18: *Keyword Search*

The search results are displayed in 4 subjects: Genes, Families, Pathways and Ontology terms. The number of results for each subject is on the left side (red circle). In the example in Figure 2.18, the search term of *mdm2* yielded results of 57 genes, 9 families, 6 pathways and 30 ontology terms.

You can simply click the numbers or the subject terms (eg., Genes) to display the results in list pages . You can also refine the search criteria by the following two ways:

- Select genome(s) from the 85 organisms in the PANTHER system. The default is all.
- Clicking the "Refine" link on the right side (blue circle), and make selections on the expanded panel (Figure 2.19).

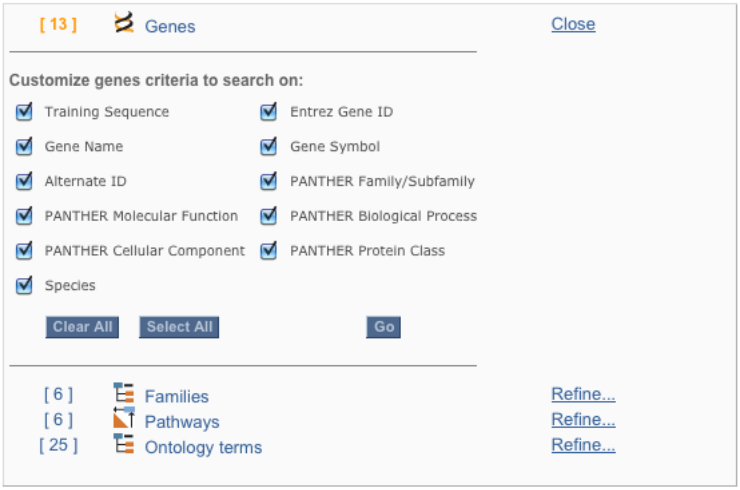


Figure 2.19: *The search criteria can be refined by clicking the* refine *link for each subject.*

Chapter 3

PANTHER Pages

PANTHER data are stored in a database in the backend. PANTHER website presents the data to the users through a number of text pages and graph/diagram pages.

The text pages include:

- Subject main page
- List pages
- Information detail pages

The graph and diagram pages include:

- Pie charts
- Pathway diagrams
- Tree Viewer

3.1 Subject main page

3.1.1 Genes and orthologs

This page can be reached from the PANTHER Data menu and then Genes & Ortholgs submenu (Figure 3.1).

PANTHER now contains the complete sets of protein coding genes for 85 different organisms, obtained from the Reference Proteome Project. This page serves as a portal for you to access information related to genes, genomes and orthologs (Figure 3.1). There are 4 links on this page that allows you to

- perform *simple keyword search* (Section 2.5);
- find *Whole Genome Pie Charts* (Section 3.4.1) from all 82 organisms;
- perform *batch ID search* (see Section 2.1 and its subSection 2.1.4.1);
- view statistics, such as, for each organism, number of total genes, number of genes in PANTHER families, number of genes with GO classifications, etc.

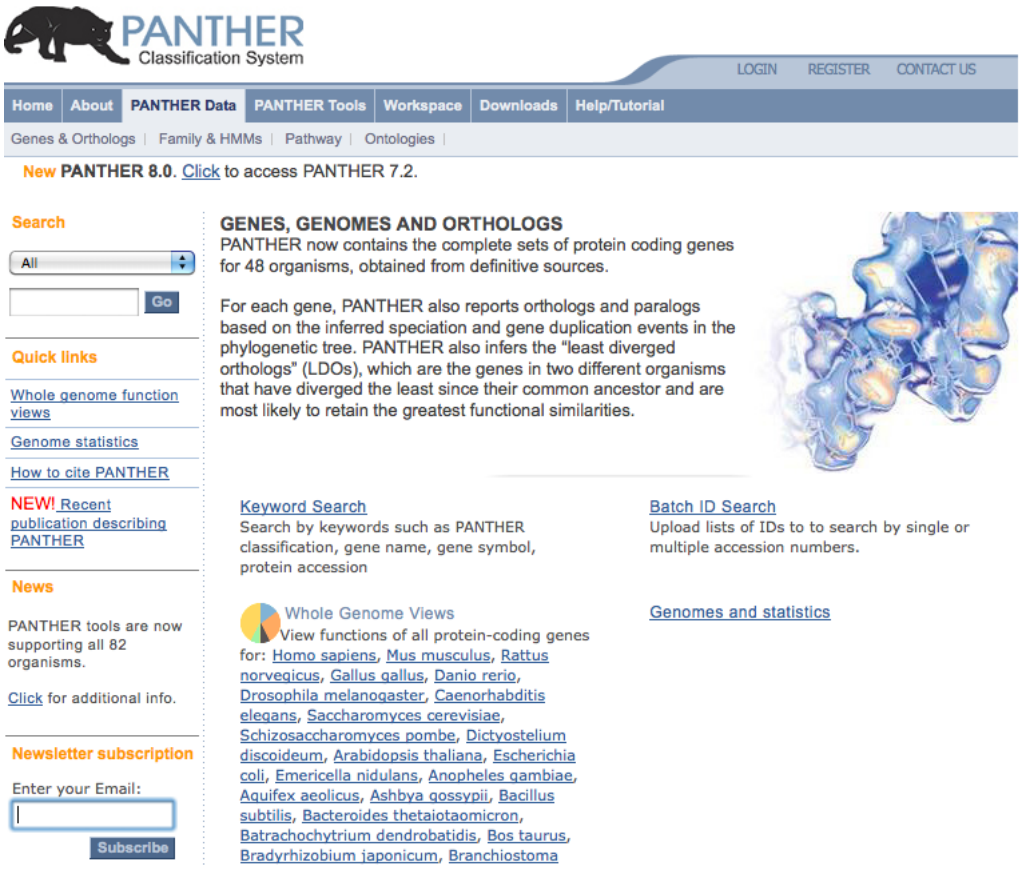

PANTHER
Classification System
LOGIN REGISTER CONTACT US
Home About PANTHER Data PANTHER Tools Workspace Downloads Help/Tutorial
Genes & Orthologs | Family & HMMs | Pathway | Ontologies |
New PANTHER 8.0. Click to access PANTHER 7.2.
Search
All
Go
Quick links
Whole genome function views
Genome statistics
How to cite PANTHER
NEW! Recent publication describing PANTHER
News
PANTHER tools are now supporting all 82 organisms.
Click for additional info.
Newsletter subscription
Enter your Email:
Subscribe
GENES, GENOMES AND ORTHOLOGS
PANTHER now contains the complete sets of protein coding genes for 48 organisms, obtained from definitive sources.
For each gene, PANTHER also reports orthologs and paralogs based on the inferred speciation and gene duplication events in the phylogenetic tree. PANTHER also infers the "least diverged orthologs" (LDOs), which are the genes in two different organisms that have diverged the least since their common ancestor and are most likely to retain the greatest functional similarities.
Keyword Search
Search by keywords such as PANTHER classification, gene name, gene symbol, protein accession
Batch ID Search
Upload lists of IDs to to search by single or multiple accession numbers.
Whole Genome Views
View functions of all protein-coding genes for: Homo sapiens, Mus musculus, Rattus norvegicus, Gallus gallus, Danio rerio, Drosophila melanogaster, Caenorhabditis elegans, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Dictyostelium discoideum, Arabidopsis thaliana, Escherichia coli, Emericella nidulans, Anopheles gambiae, Aquifex aeolicus, Ashbya gossypii, Bacillus subtilis, Bacteroides thetaiotaomicron, Batrachochytrium dendrobatidis, Bos taurus, Bradyrhizobium japonicum, Branchiostoma
Genomes and statistics

Figure 3.1: *Gene and orthologs home page.*

3.1.2 Family and HMMs

This page can be reached from the PANTHER Data menu and then Family & HMMs submenu (Figure 3.2).

PANTHER version 9.0 contains 7181 protein families, each with a phylogenetic tree relating modern-day genes in 85 organisms. Phylogenetic trees also infer the evolutionary events that occurred, such as speciation and gene duplication events. Expert biologists have divided each family into subfamilies, which are generally orthologous groups but may also contain recently duplicated paralogs. Each family and subfamily is also represented as a hidden Markov model (HMM), which can be used to classify new sequences to an existing subfamily.

This page serves as a portal for users to access information related to the trees and HMMs (Figure 3.2). There are 4 links on this page that allows you to

- search for PANTHER families, trees and HMMs. This basically takes you to the keyword search page (Section 2.5);
- browse families and subfamilies by classifications, which takes you to the prowler page (Section 2.2);
- score a new sequence against PANTHER HMMs, which takes you to the HMM score page (Section 2.3);
- download the PANTHER scoring tool to score multiple sequences against PANTHER HMMs (Section 4.1)

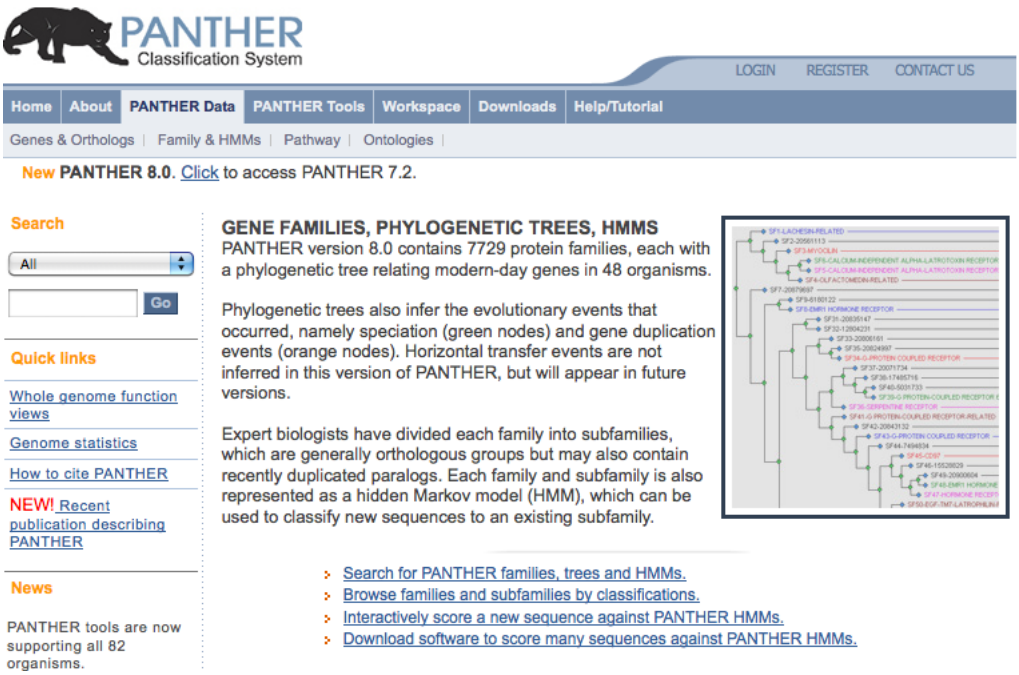

PANTHER
Classification System
LOGIN REGISTER CONTACT US
Home About PANTHER Data PANTHER Tools Workspace Downloads Help/Tutorial
Genes & Orthologs | Family & HMMs | Pathway | Ontologies |
New PANTHER 8.0. Click to access PANTHER 7.2.
Search
All
Go
Quick links
Whole genome function views
Genome statistics
How to cite PANTHER
NEW! Recent publication describing PANTHER
News
PANTHER tools are now supporting all 82 organisms.
GENE FAMILIES, PHYLOGENETIC TREES, HMMS
PANTHER version 8.0 contains 7729 protein families, each with a phylogenetic tree relating modern-day genes in 48 organisms.
Phylogenetic trees also infer the evolutionary events that occurred, namely speciation (green nodes) and gene duplication events (orange nodes). Horizontal transfer events are not inferred in this version of PANTHER, but will appear in future versions.
Expert biologists have divided each family into subfamilies, which are generally orthologous groups but may also contain recently duplicated paralogs. Each family and subfamily is also represented as a hidden Markov model (HMM), which can be used to classify new sequences to an existing subfamily.
Search for PANTHER families, trees and HMMs.
Browse families and subfamilies by classifications.
Interactively score a new sequence against PANTHER HMMs.
Download software to score many sequences against PANTHER HMMs.

Figure 3.2: *Trees and HMMs home page.*

3.1.3 Pathways

The current version of PANTHER Pathway consists of 175 pathways, each with subfamilies and protein sequences mapped to individual pathway molecule class (aka component). All pathways are stored in both SBML and BioPAX format. Pathway diagrams are interactive and include tools for visualizing gene expression data in the context of the diagrams.

This page provides a port for users to access information related to PANTHER pathways (Figure 3.3), and provides the following links.

- Community Curation - this link doesn't work since the site is under modification at the moment.
- Browse pathways - this takes you to the prowler page (section 2.2).
- Search pathways - this takes you to the advanced keyword search page (section 2.5).
- Pathway resources - links to some useful pathway databases, tools and resources.
- List of pathways - lists all 175 PANTHER pathways, and links to view the interactive pathway diagram, download SBML and BioPAX files.
- Download pathway data - it links to the ftp site for all pathway data download (Section 4.1).

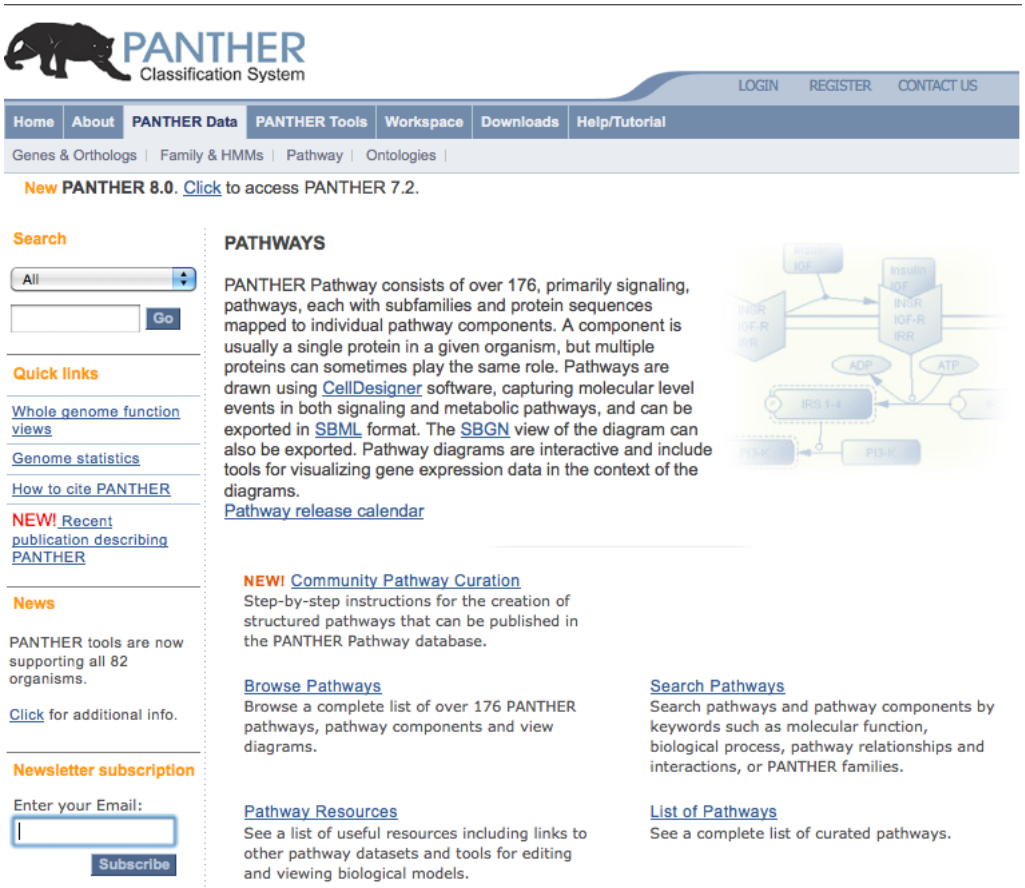

PANTHER
Classification System
LOGIN
REGISTER
CONTACT US
Home
About
PANTHER Data
PANTHER Tools
Workspace
Downloads
Help/Tutorial
Genes & Orthologs | Family & HMMs | Pathway | Ontologies |
New PANTHER 8.0. Click to access PANTHER 7.2.
Search
All
Go
Quick links
Whole genome function views
Genome statistics
How to cite PANTHER
NEW! Recent publication describing PANTHER
News
PANTHER tools are now supporting all 82 organisms.
Click for additional info.
Newsletter subscription
Enter your Email:
Subscribe
PATHWAYS
PANTHER Pathway consists of over 176, primarily signaling, pathways, each with subfamilies and protein sequences mapped to individual pathway components. A component is usually a single protein in a given organism, but multiple proteins can sometimes play the same role. Pathways are drawn using CellDesigner software, capturing molecular level events in both signaling and metabolic pathways, and can be exported in SBML format. The SBGN view of the diagram can also be exported. Pathway diagrams are interactive and include tools for visualizing gene expression data in the context of the diagrams.
Pathway release calendar
NEW! Community Pathway Curation
Step-by-step instructions for the creation of structured pathways that can be published in the PANTHER Pathway database.
Browse Pathways
Browse a complete list of over 176 PANTHER pathways, pathway components and view diagrams.
Search Pathways
Search pathways and pathway components by keywords such as molecular function, biological process, pathway relationships and interactions, or PANTHER families.
Pathway Resources
See a list of useful resources including links to other pathway datasets and tools for editing and viewing biological models.
List of Pathways
See a complete list of curated pathways.

Figure 3.3: *PANTHER Pathway home page.*

3.1.4 Ontologies

This page provides a brief description of the ontolgoies used in PANTHER. It also provides links to browse in prowler (Section 2.2) and download the ontologies.

3.1.5 PANTHER Tools

PANTHER System is not only a database, but also a data analysis platform that allows users to perform more advanced tasks, and analyze research results using this integrated system. This page provides links to the following tools:

- Gene expression data analysis tools (see detail in Section 2.1)
- Evolutionary analysis of coding SNPs (see detail in Section 2.4)
- PANTHER scoring (see detail in Section 2.3)
- Downloads (see Section 4.1)

3.1.6 Workspace

The Workspace is a unique feature in PANTHER that allows users to store the gene lists that they generate for future analysis. Although users do not have to register to use the PANTHER system, registration is required in order to user the workspace. Registration is free. This webpage allows registered users to login, or retrieve password, and for non-registered users to register.

A more detailed description of using *Workspace* can be found in Section 4.2.

3.2 List pages

3.2.1 Gene list page

Thge gene list page can be generated from the gene list analysis tool (functional classification option), prowler, keyword search, pathway diagram, and other list and detail pages.
The gene list page contains the following information (Figure 3.4).

- **Gene ID** – This is the identifier for genes in the PANTHER library. The format is as follows: organism|gene database source=gene id|protein database source=protein id. For example, HUMAN|ENSEMBL=ENSG00000111262|UniProtKB=Q09470 is a human sequence, the gene sequence is from ENSEMBL with id ENSG00000111262, and the protein sequence is from UniProt with id Q09470. Clicking on the ID will lead to the Gene Detail page (section 3.3.1).
- **Gene Name/Gene Symbol** – The Entrez gene definition and gene symbol. Clicking on the gene symbol will lead to the Gene Detail page (section 3.3.1). Please note that some of the genes do not have the information in this column.
- **PANTHER Family/Subfamily** – The family or subfamily name of the PANTHER model where the sequence is in. Clicking on the name will lead to the PANTHER family or subfamily detail page (section 3.3.2).
- **PANTHER Protein Class** – This is a PANTHER Index terms describing protein classes. Clicking on the term will lead to the term detail page (section 3.3.3).
- **Species** – The organism of the gene.

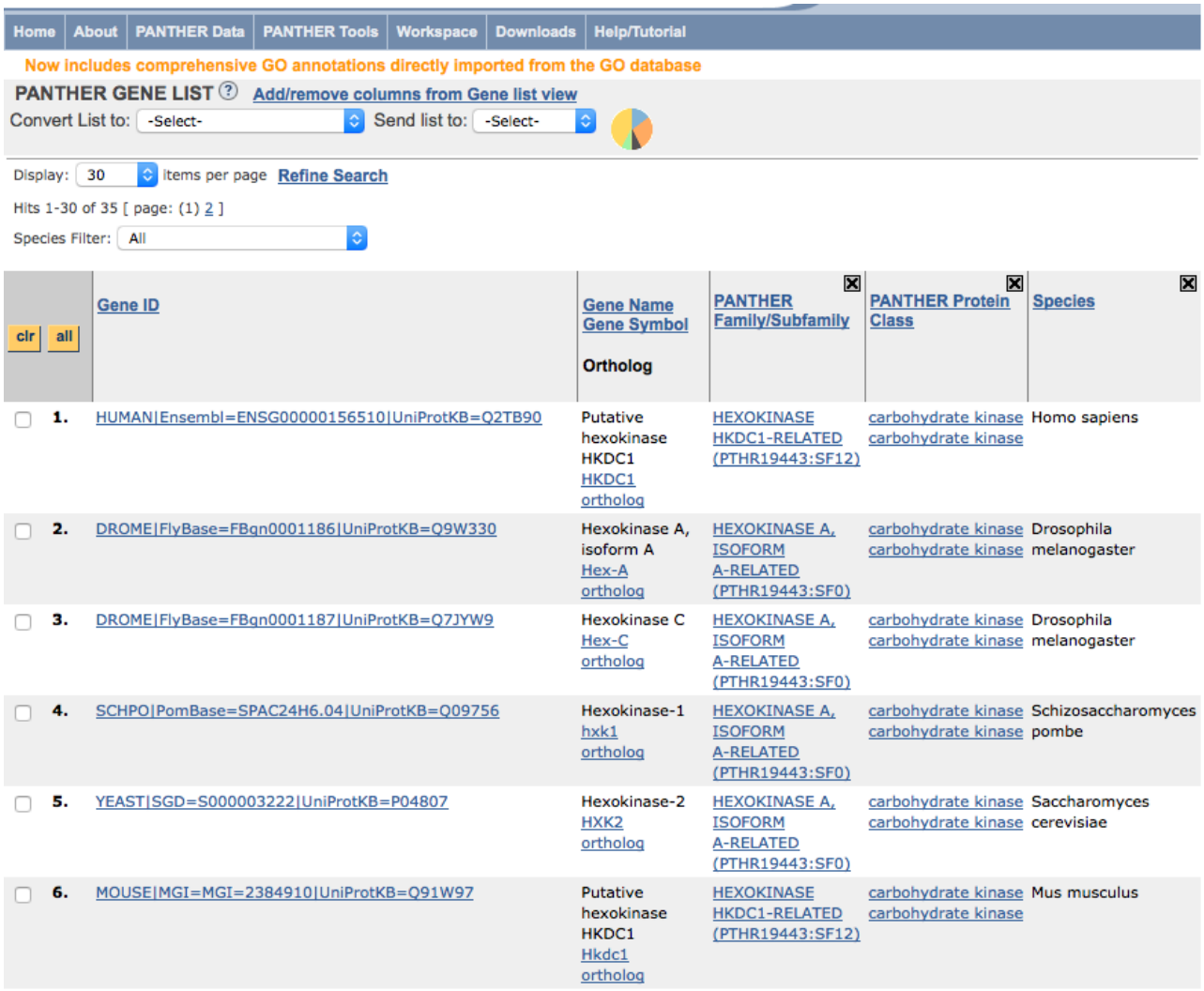

Home
About
PANTHER Data
PANTHER Tools
Workspace
Downloads
Help/Tutorial
Now includes comprehensive GO annotations directly imported from the GO database
PANTHER GENE LIST
Add/remove columns from Gene list view
Convert List to: -Select-
Send list to: -Select-
Display: 30 items per page
Refine Search
Hits 1-30 of 35 [page: (1) 2]
Species Filter: All
clr
all
Gene ID
Gene Name
Gene Symbol
Ortholog
PANTHER Family/Subfamily
PANTHER Protein Class
Species
1. HUMAN|Ensembl=ENSG00000156510|UniProtKB=Q2TB90 Putative hexokinase HKDC1 HKDC1 ortholog HEXOKINASE HKDC1-RELATED (PTHR19443:SF12) carbohydrate kinase carbohydrate kinase Homo sapiens
2. DROME|FlyBase=FBgn0001186|UniProtKB=Q9W330 Hexokinase A, isoform A Hex-A ortholog HEXOKINASE A, ISOFORM A-RELATED (PTHR19443:SF0) carbohydrate kinase carbohydrate kinase Drosophila melanogaster
3. DROME|FlyBase=FBgn0001187|UniProtKB=Q7JYW9 Hexokinase C Hex-C ortholog HEXOKINASE A, ISOFORM A-RELATED (PTHR19443:SF0) carbohydrate kinase carbohydrate kinase Drosophila melanogaster
4. SCHPO|PomBase=SPAC24H6.04|UniProtKB=Q09756 Hexokinase-1 hxk1 ortholog HEXOKINASE A, ISOFORM A-RELATED (PTHR19443:SF0) carbohydrate kinase carbohydrate kinase Schizosaccharomyces pombe
5. YEAST|SGD=S000003222|UniProtKB=P04807 Hexokinase-2 HXK2 ortholog HEXOKINASE A, ISOFORM A-RELATED (PTHR19443:SF0) carbohydrate kinase carbohydrate kinase Saccharomyces cerevisiae
6. MOUSE|MGI=MGI=2384910|UniProtKB=Q91W97 Putative hexokinase HKDC1 Hkdc1 ortholog HEXOKINASE HKDC1-RELATED (PTHR19443:SF12) carbohydrate kinase carbohydrate kinase Mus musculus

Figure 3.4: *Gene list page.*

Tip: Manipulate your list

Here is a description of what you can do to your list in any list pages in this section (3.2).

- Sort the list – You can always sort the list by clicking on any of the underlined column names. A yellow triangle appears in front of the column name that you choose to sort. The orientation of the triangle indicates the sort is ascending or descending.
- Customize columns: You can click on the "x" button next to the column names to collapse the column. You can also click the "Add/remove column from Gene list view" to add additional columns, such as PANTHER GO-slim, full GO or PANTHER pathway annotations.
- Converting a list to another list type. Select the genes you want to convert by clicking the checkboxes. The default is for all genes in the list.
  - Click on the pull-down menu after *Convert list to*. The current list type is shown in the box.
  - Select the new list type from the pull-down menu (Figure 3.5).

– Each primary ID (first column of the list) is used to return the selected data type. Note that the mapping between different types is not necessarily one-to-one (e.g. a subfamily can map to more than one associated transcript).

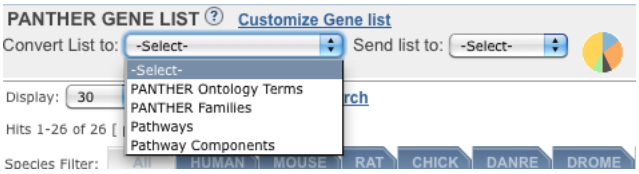


Figure 3.5: *Options to convert a list from one type to another.*

- Saving a list to your workspace. Select the genes you want to save by clicking the check-boxes. The default is for all genes in the list. You can select one of the followings from the pull-down menu as the destination (Figure 3.6).
  - Workspace – You need to register to save data to the workspace. The registration is free. When you make this selection, a pop-up window will ask you to name the list and add any comments. The name and comments can be edited at any time in the future from the Workspace page. Once the gene list is now at the site, it can be returned to at any time. Only the IDs are stored, and they are mapped to the internal PANTHER gene ids, so when you access a list in the future, all information will be updated and current.
  - Exporting a list to a file – The list will be exported as a tab-delimited file. You can now import the file into Excel or perform any post-processing you wish.
  - View the list as text on the website.
- Use the pie chart view by clicking the colorful pie chart icon. See Pie Charts section (3.4.1) for details.

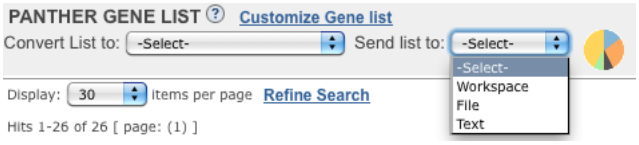


Figure 3.6: *Export list.*

3.2.2 Family/subfamily list page

The family/subfamily list page can be generated from prowler, keyword search, pathway diagram, and other list and detail pages.
This page provides a list of families and/or subfamilies with the following related information (Figure 3.7).

- **Family ID** – In this column, it usually lists either family or subfamily IDs. Click the ID will take you to the *Family Detail page* (see section 3.3.2).
- **Family Name** – This column lists the names of families or subfamilies. Click the name will take you to the *Family Detail page* (see section 3.3.2).

- **Genes** – This column lists the number of Genes in the family or subfamily. These are also the sequences used to build the model. Clicking on the number will take you to the *Gene List page* (see section 3.2.1).
- **PANTHER Protein Class** – This is a PANTHER Index terms describing protein classes. You can click on each term to view more information in the *Ontology Term Detail page* (see section 3.3.3).
- **Interpro** – If the PANTHER family or subfamily is integrated into Interpro, the Interpro accession is listed here. Clicking on the Interpro accession will lead you to the Interpro website.

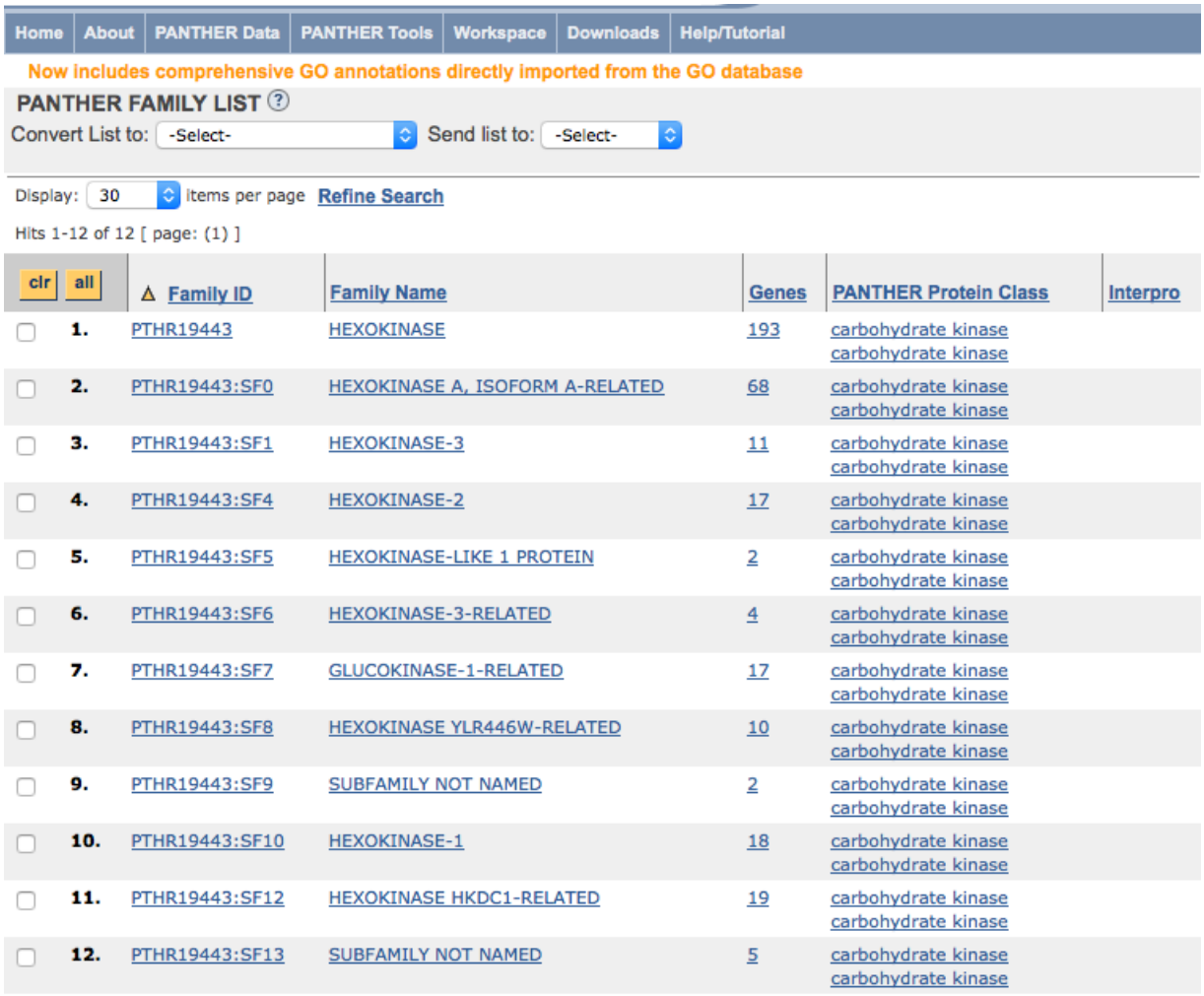

Home
About
PANTHER Data
PANTHER Tools
Workspace
Downloads
Help/Tutorial
Now includes comprehensive GO annotations directly imported from the GO database
PANTHER FAMILY LIST
Convert List to: -Select-
Send list to: -Select-
Display: 30 items per page Refine Search
Hits 1-12 of 12 [page: (1)]
clr all
Family ID
Family Name
Genes
PANTHER Protein Class
Interpro
1. PTHR19443 HEXOKINASE 193 carbohydrate kinase carbohydrate kinase
2. PTHR19443:SF0 HEXOKINASE A, ISOFORM A-RELATED 68 carbohydrate kinase carbohydrate kinase
3. PTHR19443:SF1 HEXOKINASE-3 11 carbohydrate kinase carbohydrate kinase
4. PTHR19443:SF4 HEXOKINASE-2 17 carbohydrate kinase carbohydrate kinase
5. PTHR19443:SF5 HEXOKINASE-LIKE 1 PROTEIN 2 carbohydrate kinase carbohydrate kinase
6. PTHR19443:SF6 HEXOKINASE-3-RELATED 4 carbohydrate kinase carbohydrate kinase
7. PTHR19443:SF7 GLUCOKINASE-1-RELATED 17 carbohydrate kinase carbohydrate kinase
8. PTHR19443:SF8 HEXOKINASE YLR446W-RELATED 10 carbohydrate kinase carbohydrate kinase
9. PTHR19443:SF9 SUBFAMILY NOT NAMED 2 carbohydrate kinase carbohydrate kinase
10. PTHR19443:SF10 HEXOKINASE-1 18 carbohydrate kinase carbohydrate kinase
11. PTHR19443:SF12 HEXOKINASE HKDC1-RELATED 19 carbohydrate kinase carbohydrate kinase
12. PTHR19443:SF13 SUBFAMILY NOT NAMED 5 carbohydrate kinase carbohydrate kinase

Figure 3.7: *Family/subfamily list page.*

You can sort the list, convert the list to a different ID type and save the list as described in Gene List section (3.2.1).

3.2.3 Pathway list page

This page provides a list of pathways with the following information (Figure 3.8).

- Pathway Accession - This column lists the unique accession of each pathway. Clicking the accession will take you to the *Pathway Diagram* (section 3.4.2).

- Pathway Name - This column lists the pathway names in the list. Clicking the accession will take you to the *Pathway Diagram* (section 3.4.2).
- Components - This column lists the number of components (or molecule classes) in each pathway. Clicking the number will take you to the pathway component list page.
- Subfamilies - This column provides the number of subfamilies of all genes that are associated to each pathway. Clicking the number will take you to the *Subfamily list page* (section 3.2.2).
- Associated Sequences - This column lists the numbers of genes that are associated to the pathway. Clicking the number will take you to the *Gene list page* (section 3.2.1).

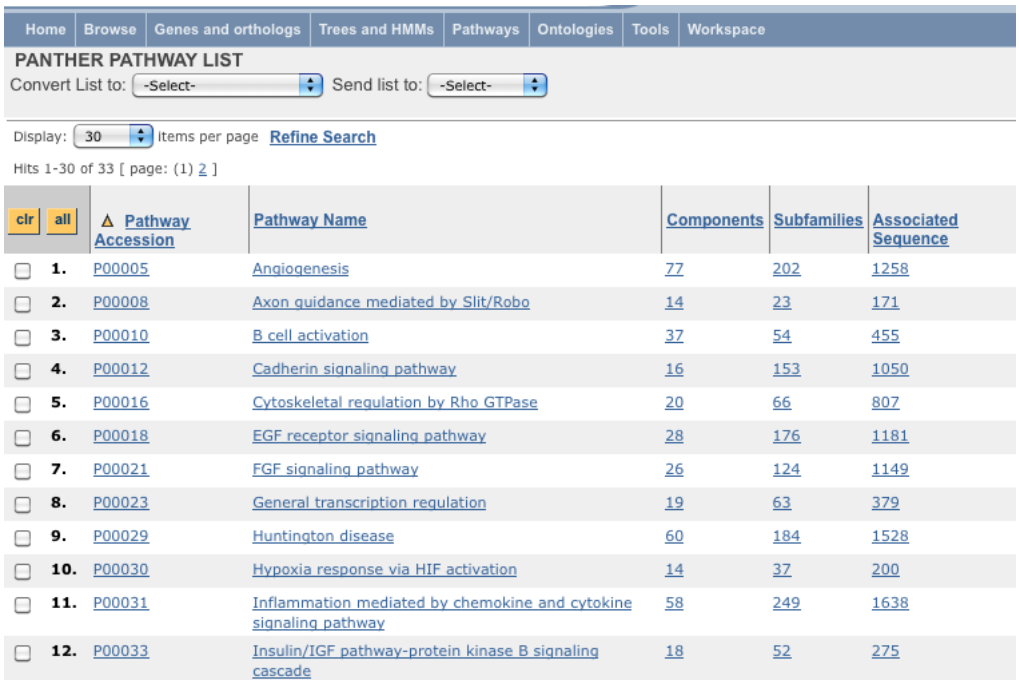

Home
Browse
Genes and orthologs
Trees and HMMs
Pathways
Ontologies
Tools
Workspace
PANTHER PATHWAY LIST
Convert List to: -Select-
Send list to: -Select-
Display: 30 items per page
Refine Search
Hits 1-30 of 33 [page: (1) 2]
clr
all
Pathway Accession
Pathway Name
Components
Subfamilies
Associated Sequence
1. P00005 Angiogenesis 77 202 1258
2. P00008 Axon guidance mediated by Slit/Robo 14 23 171
3. P00010 B cell activation 37 54 455
4. P00012 Cadherin signaling pathway 16 153 1050
5. P00016 Cytoskeletal regulation by Rho GTPase 20 66 807
6. P00018 EGF receptor signaling pathway 28 176 1181
7. P00021 FGF signaling pathway 26 124 1149
8. P00023 General transcription regulation 19 63 379
9. P00029 Huntington disease 60 184 1528
10. P00030 Hypoxia response via HIF activation 14 37 200
11. P00031 Inflammation mediated by chemokine and cytokine signaling pathway 58 249 1638
12. P00033 Insulin/IGF pathway-protein kinase B signaling cascade 18 52 275

Figure 3.8: *Pathway list page.*

You can sort the list, convert the list to a different ID type and save the list as described in Gene List section (3.2.1).

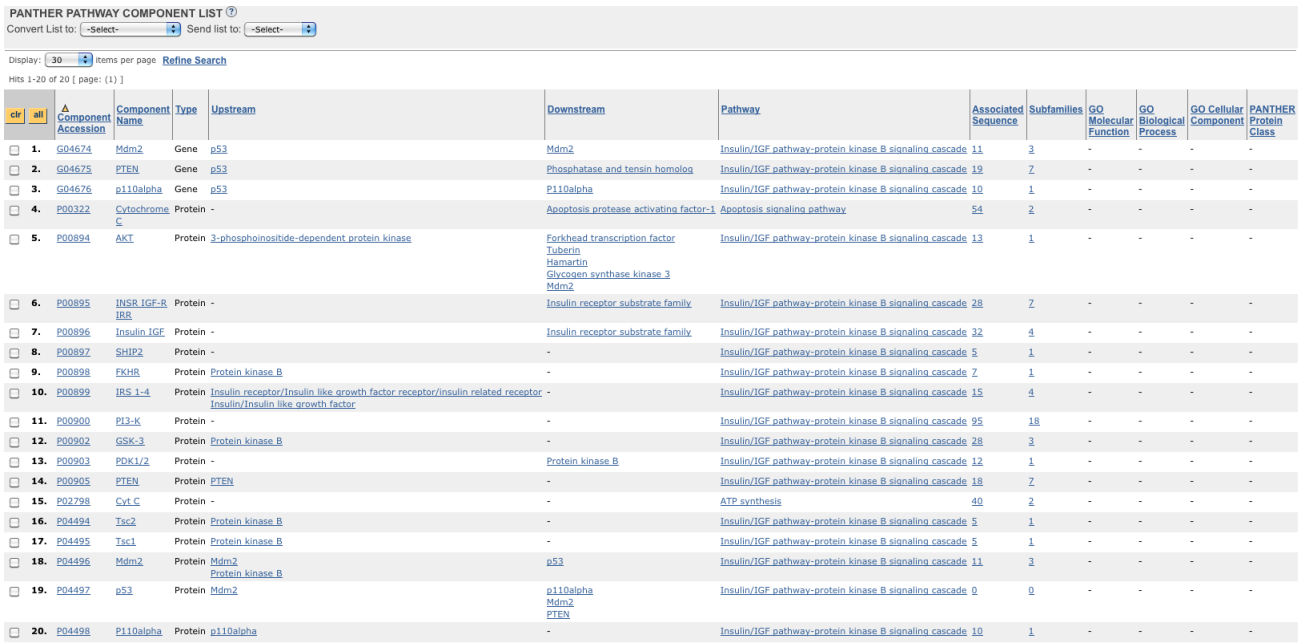
3.2.4 Pathway component list page

A pathway component (also called molecule class) represents a group of homologous proteins across various organisms that participate in the same specific biochemical reactions within the pathway. A pathway component is always associated with training sequences within the PANTHER libraries, thus, it is linked directly or indirectly to the PANTHER family/subfamily statistical models.

This page lists pathway components with the following information (Figure 3.9):

- Component Accession - The accession assigned to each pathway component. It is usually a 5 digit number proceeded by a letter to indicate the type of the component in the pathway (G for DNA or RNA, P for protein). Clicking it will lead to the *Pathway molecule class detail page* (section 3.3.5).

- Component Name - The display name of the component. Clicking it lead to the *Pathway molecule class detail page* (section 3.3.5).
- Type - Indicates the type of the component in the pathway, eg., gene, protein.
- Upstream and Downstream - These two columns list the pathway component(s) immediately upstream or downstream of the component. Upstream is defined as the component that controls or modulates the reaction that the current component is involved in. Downstream component(s) are those involved in the reaction that is controlled or modulated by the current component. Clicking it lead to the *Pathway molecule class detail page* (section 3.3.5).
- Pathway - The pathway that the component belongs to. Clicking it lead to the *Pathway molecule diagram page* (section 3.4.2).
- Associated Sequences - This column lists the numbers of genes that are associated to this pathway component. Clicking the number will take you to the *Gene list page* (section 3.2.1).
- Subfamilies - This column provides the number of subfamilies of all genes that are associated to the pathway component. Clicking the number will take you to the *Subfamily list page* (section 3.2.2).
- GO classifications and PANTHER protein class - These 4 columns lists all the categories classified to the genes that are associated to the pathway component. Clicking on an ontology terms will lead to the *Ontology Term Details page* (section 3.3.3)..



PANTHER PATHWAY COMPONENT LIST

Convert List to: -Select- Send list to: -Select-

Display: 30 items per page Refine Search

Hits 1-20 of 20 [page: (1)]

clr all		Component Accession	Component Name	Type	Upstream	Downstream	Pathway	Associated Sequence	Subfamilies	GO Molecular Function	GO Biological Process	GO Cellular Component	PANTHER Protein Class
	1.	G04674	Mdm2	Gene	p53	Mdm2	Insulin/IGF pathway-protein kinase B signaling cascade	11	3	-	-	-	-
	2.	G04675	PTEN	Gene	p53	Phosphatase and tensin homolog	Insulin/IGF pathway-protein kinase B signaling cascade	19	7	-	-	-	-
	3.	G04676	p110alpha	Gene	p53	P110alpha	Insulin/IGF pathway-protein kinase B signaling cascade	10	1	-	-	-	-
	4.	P00322	Cytochrome C	Protein	-	Apoptosis protease activating factor-1	Apoptosis signaling pathway	54	2	-	-	-	-
	5.	P00894	AKT	Protein	3-phosphoinositide-dependent protein kinase	Forkhead transcription factor Tuberin Hamartin Glycogen synthase kinase 3 Mdm2	Insulin/IGF pathway-protein kinase B signaling cascade	13	1	-	-	-	-
	6.	P00895	INSR IGF-R IRR	Protein	-	Insulin receptor substrate family	Insulin/IGF pathway-protein kinase B signaling cascade	28	7	-	-	-	-
	7.	P00896	Insulin IGF	Protein	-	Insulin receptor substrate family	Insulin/IGF pathway-protein kinase B signaling cascade	32	4	-	-	-	-
	8.	P00897	SHIP2	Protein	-	-	Insulin/IGF pathway-protein kinase B signaling cascade	5	1	-	-	-	-
	9.	P00898	FKHR	Protein	Protein kinase B	-	Insulin/IGF pathway-protein kinase B signaling cascade	7	1	-	-	-	-
	10.	P00899	IRS 1-4	Protein	Insulin receptor/Insulin like growth factor receptor/insulin related receptor Insulin/Insulin like growth factor	-	Insulin/IGF pathway-protein kinase B signaling cascade	15	4	-	-	-	-
	11.	P00900	PI3-K	Protein	-	-	Insulin/IGF pathway-protein kinase B signaling cascade	95	18	-	-	-	-
	12.	P00902	GSK-3	Protein	Protein kinase B	-	Insulin/IGF pathway-protein kinase B signaling cascade	28	3	-	-	-	-
	13.	P00903	PDK1/2	Protein	-	Protein kinase B	Insulin/IGF pathway-protein kinase B signaling cascade	12	1	-	-	-	-
	14.	P00905	PTEN	Protein	PTEN	-	Insulin/IGF pathway-protein kinase B signaling cascade	18	7	-	-	-	-
	15.	P02798	Cyt C	Protein	-	-	ATP synthesis	40	2	-	-	-	-
	16.	P04494	Tsc2	Protein	Protein kinase B	-	Insulin/IGF pathway-protein kinase B signaling cascade	5	1	-	-	-	-
	17.	P04495	Tsc1	Protein	Protein kinase B	-	Insulin/IGF pathway-protein kinase B signaling cascade	5	1	-	-	-	-
	18.	P04496	Mdm2	Protein	Mdm2 Protein kinase B	p53	Insulin/IGF pathway-protein kinase B signaling cascade	11	3	-	-	-	-
	19.	P04497	p53	Protein	Mdm2	p110alpha Mdm2 PTEN	Insulin/IGF pathway-protein kinase B signaling cascade	0	0	-	-	-	-
	20.	P04498	P110alpha	Protein	p110alpha	-	Insulin/IGF pathway-protein kinase B signaling cascade	10	1	-	-	-	-

Figure 3.9: *Pathway component list page.*

You can sort the list, convert the list to a different ID type and save the list as described in Gene List section (3.2.1).

3.2.5 Ontology term list page

This page provides a list of PANTHER ontology terms, including PANTHER GO Slim and PANTHER Protein Class, with the following related information.

- Category ID - A GO id (for PANTHER GO Slim) or a PANTHER Protein Class id, which is a 5 digit number preceded by letters PC. Clicking on the ID will lead to the *Ontology Term Detail page* (section 3.3.3).
- Name - Category names.
- Parent: The names of all parent categories. A parent category refers to a more general category of the current one, e.g., Receptor is a parent category of Protein kinase receptor. For GO terms, the parent is not the parent category in the full Gene Ontology, but the parent GO terms in the PANTHER GO Slim. Clicking on the category name will lead to the Category Detail page.
- Child: The names of all child categories. The names of all child categories. A child category refers to a more specific category of the current one, e.g., Serine/threonine protein kinase receptor is a child category of Protein kinase receptor. For GO terms, the child is not the parent category in the full Gene Ontology, but the child GO terms in the PANTHER GO Slim. Clicking on the category name will lead to the Category Detail page.
- Families and Subfamilies: These 2 columns list the number of families and subfamilies that are classified by the category.

The yellow triangle in front of the column name indicates that the table is currently sorted by the column.

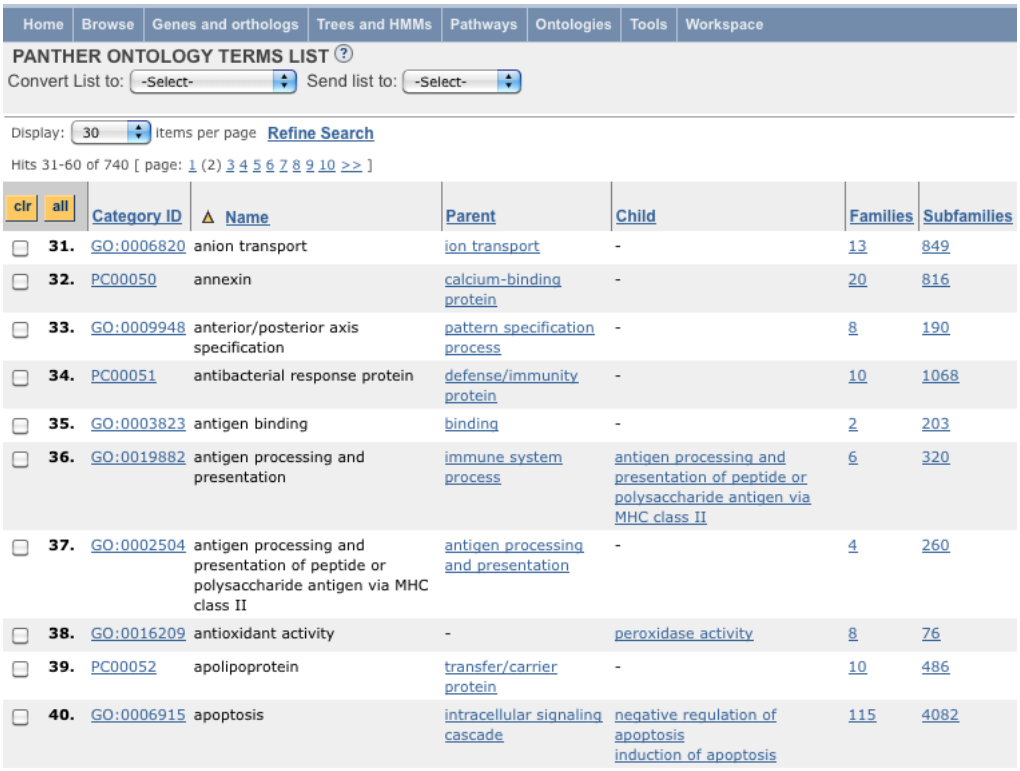

Home
Browse
Genes and orthologs
Trees and HMMs
Pathways
Ontologies
Tools
Workspace
PANTHER ONTOLOGY TERMS LIST
Convert List to: -Select-
Send list to: -Select-
Display: 30 items per page Refine Search
Hits 31-60 of 740 [page: 1 (2) 3 4 5 6 7 8 9 10 >>]
clr all
Category ID
Name
Parent
Child
Families
Subfamilies
31. GO:0006820 anion transport ion transport - 13 849
32. PC00050 annexin calcium-binding protein - 20 816
33. GO:0009948 anterior/posterior axis specification pattern specification process - 8 190
34. PC00051 antibacterial response protein defense/immunity protein - 10 1068
35. GO:0003823 antigen binding binding - 2 203
36. GO:0019882 antigen processing and presentation immune system process antigen processing and presentation of peptide or polysaccharide antigen via MHC class II 6 320
37. GO:0002504 antigen processing and presentation of peptide or polysaccharide antigen via MHC class II antigen processing and presentation - 4 260
38. GO:0016209 antioxidant activity - peroxidase activity 8 76
39. PC00052 apolipoprotein transfer/carrier protein - 10 486
40. GO:0006915 apoptosis intracellular signaling cascade negative regulation of apoptosis induction of apoptosis 115 4082

Figure 3.10: *Ontology term list page.*

You can sort the list, convert the list to a different ID type and save the list as described in Gene List section (3.2.1).

3.3 Information detail pages

3.3.1 Gene detail page

This page provides detailed information about a gene. The page contains 3 general sections.

1. **General information about the gene (Figure 3.11)**

   Gene ID From source database, either a model organism database (e.g. FlyBase), Ensembl, or Entrez Gene.

   Protein ID A UniProt (preferentially SwissProt) ID whenever possible, this is the "representative" protein sequence for the gene.

   Gene Name From source database, together with the Gene ID.

   Gene Symbol(s) Entrez gene symbol, if available.

   Organism The organism of the gene.

   Alternative IDs Identifiers in other databases for this gene or a protein product of the gene. This is mapped based on the UniProt ID Mapping tool.

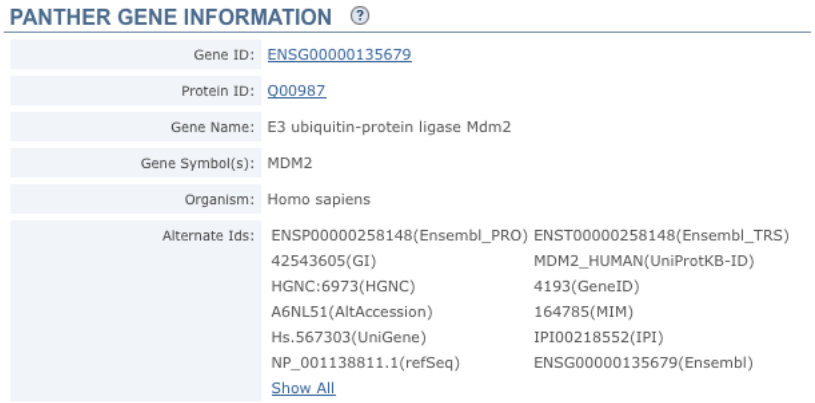


Figure 3.11: *Gene information section of the gene detail page.*

2. **PANTHER classification of the gene (Figure 3.12)**

   PANTHER family The family to which the gene belongs; there is a phylogenetic tree for this family. This link will bring up the PANTHER family page.

   PANTHER subfamily The subfamily to which the gene belongs; the subfamily is annotated on the phylogenetic tree. This link will bring up the PANTHER subfamily page.

   GO molecular function, biological process, cellular component These are Gene Ontology terms describing the function of the gene product.

   PANTHER Protein Class This is a PANTHER Index terms describing protein classes.

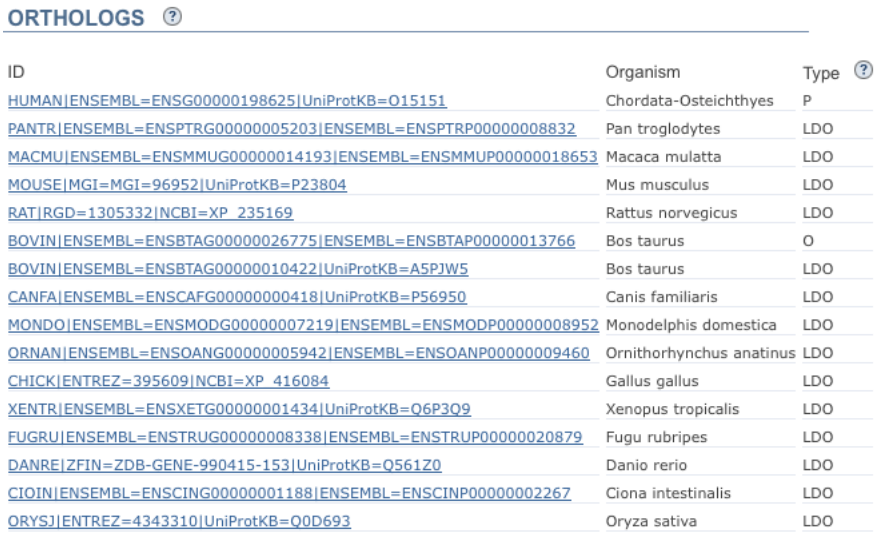
   Pathway This links to a diagram of the pathway(s) a gene product participates in.

Figure 3.12: *Gene classification section of the gene detail page.*

3. **Orthologs (Figure 3.13)**

   This section lists all the orthologs of this gene.

   - "LDO'' means least diverged ortholog, while other, more diverged orthologs (if any) are marked "O". Paralogs are marked "P".
   - For orthologs, the species column gives the modern-day species in which the ortholog is found. For paralogs, the species column gives the two speciation events between which the duplication occurred that generated the paralogous genes. "ND" means "not determined". Thus different paralogs can be distinguished by how long ago the relevant duplications occurred.
   - Orthologs are genes that can be traced to the same gene in the genome of their most recent common ancestor species. Paralogs are genes that are traced to related, but distinct, genes in the genome of their most recent common ancestor species.
   - The LDO can be loosely interpreted as the most nearly "equivalent" gene in another organism. If there were no gene duplication events following the speciation event that separated two species, there is only a single ortholog in each genome and they are therefore marked as "LDO." If there was a gene duplication event following speciation, there may be more than one ortholog and the least diverged ortholog pair is marked "LDO," while other ortholog pairs are marked "O."



ORTHOLOGS

ID	Organism	Type
HUMAN|ENSEMBL=ENSG00000198625|UniProtKB=O15151	Chordata-Osteichthyes	P
PANTR|ENSEMBL=ENSPTRG00000005203|ENSEMBL=ENSPTRP00000008832	Pan troglodytes	LDO
MACMU|ENSEMBL=ENSMMUG00000014193|ENSEMBL=ENSMMUP00000018653	Macaca mulatta	LDO
MOUSE|MGI=MGI=96952|UniProtKB=P23804	Mus musculus	LDO
RAT|RGD=1305332|NCBI=XP_235169	Rattus norvegicus	LDO
BOVIN|ENSEMBL=ENSBTAG00000026775|ENSEMBL=ENSBTAP00000013766	Bos taurus	O
BOVIN|ENSEMBL=ENSBTAG00000010422|UniProtKB=A5PJW5	Bos taurus	LDO
CANFA|ENSEMBL=ENSCAFG00000000418|UniProtKB=P56950	Canis familiaris	LDO
MONDO|ENSEMBL=ENSMODG00000007219|ENSEMBL=ENSMODP00000008952	Monodelphis domestica	LDO
ORNAN|ENSEMBL=ENSOANG00000005942|ENSEMBL=ENSOANP00000009460	Ornithorhynchus anatinus	LDO
CHICK|ENTREZ=395609|NCBI=XP_416084	Gallus gallus	LDO
XENTR|ENSEMBL=ENSXETG00000001434|UniProtKB=Q6P3Q9	Xenopus tropicalis	LDO
FUGRU|ENSEMBL=ENSTRUG00000008338|ENSEMBL=ENSTRUP00000020879	Fugu rubripes	LDO
DANRE|ZFIN=ZDB-GENE-990415-153|UniProtKB=Q561Z0	Danio rerio	LDO
CIOIN|ENSEMBL=ENSCING00000001188|ENSEMBL=ENSCINP00000002267	Ciona intestinalis	LDO
ORYSJ|ENTREZ=4343310|UniProtKB=Q0D693	Oryza sativa	LDO

Figure 3.13: *Ortholog table of the gene detail page.*

3.3.2 Family/subfamily detail page

This page provides details of a PANTHER family or subfamily model. It contains 2 sections.

1. Family/subfamily information

   The pages for family and subfamily information are slightly different. The family detail page contains the following information (Figure 3.14):

   - Family: The name of the family model. Family ID is in parentheses (e.g., PTHR10000).
   - Subfamilies: The number of subfamilies under the family. Clicking on the number will lead to the Subfamily List page (section 3.2.2).
   - PANTHER Links: Two links will lead to the Tree and MSA pages (section 3.4.3).
   - GO molecular function, biological process, cellular component: These are Gene Ontology terms from PANTHER GO Slim describing the function of the gene product.
   - PANTHER Protein Class: This is a PANTHER Index term describing protein classes.
   - PANTHER Pathway: This links to a diagram of the pathway(s) in which a gene product participates.
   - Genes: The number of genes that were used as training sequences to build the family model. Clicking on the number will lead to the Gene List Page (section 3.2.1).
   - HMM Length: The number of position in the hidden-Markov model.
   - Downloads: You can download the family HMM built with either HMMER or SAM method.

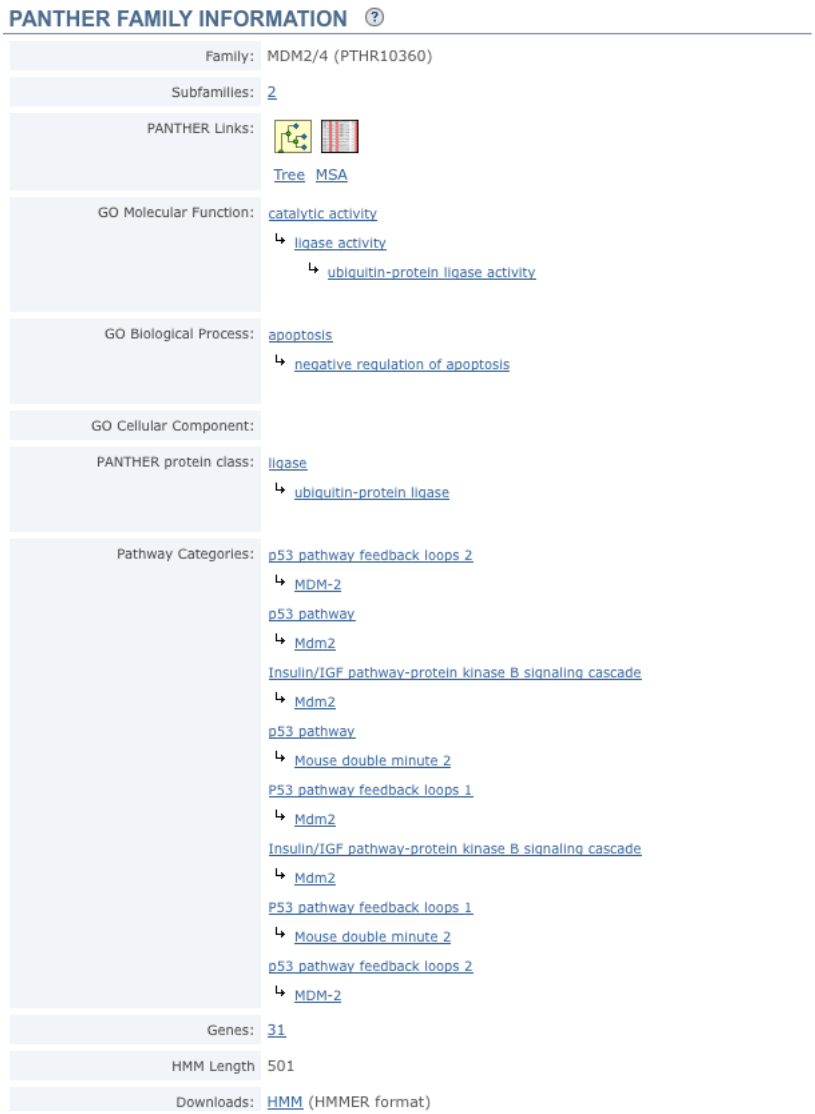


Figure 3.14: *Family information section of the family detail page.*

The subfamily detail page contains similar information as the family detail page except for the first two items (Figure 3.15):

- Subfamily: The name of the subfamily and the subfamily ID (eg., PTHR10000:SF1)
- Family: The family name that the subfamily belongs to. Clicking the name will lead you to the family detail page

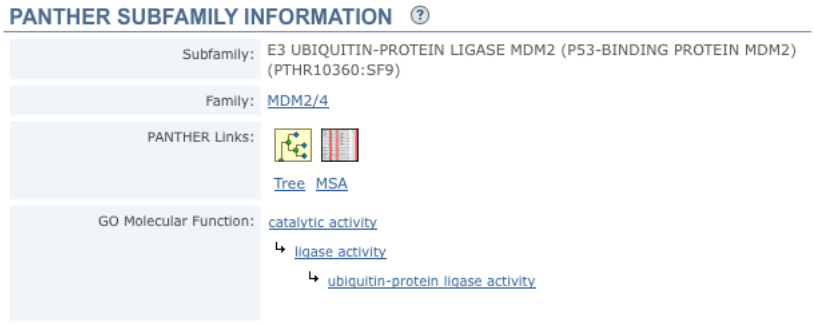
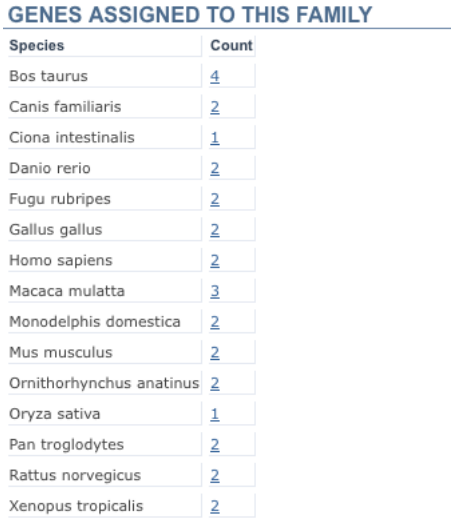


Figure 3.15: *Subfamily information section of the family detail page.*

2. Genes Assigned To This Family: A table listing the number of genes from various species that hit this model. Clicking on the number will lead to the Gene List page (section 3.2.1).



GENES ASSIGNED TO THIS FAMILY

Species	Count
Bos taurus	4
Canis familiaris	2
Ciona intestinalis	1
Danio rerio	2
Fugu rubripes	2
Gallus gallus	2
Homo sapiens	2
Macaca mulatta	3
Monodelphis domestica	2
Mus musculus	2
Ornithorhynchus anatinus	2
Oryza sativa	1
Pan troglodytes	2
Rattus norvegicus	2
Xenopus tropicalis	2

Figure 3.16: *List of genes assigned to the family in the family detail page.*

3.3.3 Ontology term detail page

This page provides the following information for a particular PANTHER GO Slim term or a PANTHER protein class category (Figure 3.17):

Name The name of the category.

Definition brief description of the category.

Class ID A GO id (for PANTHER GO Slim) or an internal tracking ID (for PANTHER Protein Class). The PANTHER id is usually a 5 digit number preceded by letters PC (for Protein Class).

Parent lineages The path of the category with all parent categories. For GO terms, the parent is not the parent category in the full Gene Ontology, but the parent GO category in the PANTHER GO Slim. Clicking the name will take you to the detail page of the ontology term.

Child The list of names of all child categories. For GO terms, the child is not the child category in the full Gene Ontology, but the child GO terms in the PANTHER GO Slim. Clicking the name will lead you to the PANTHER detail page of the child category.

Number of subfamilies/Families The number of subfamilies and families that are classified by the category. Clicking on the number will lead to the Family/Subfamily List page (section 3.3.2).

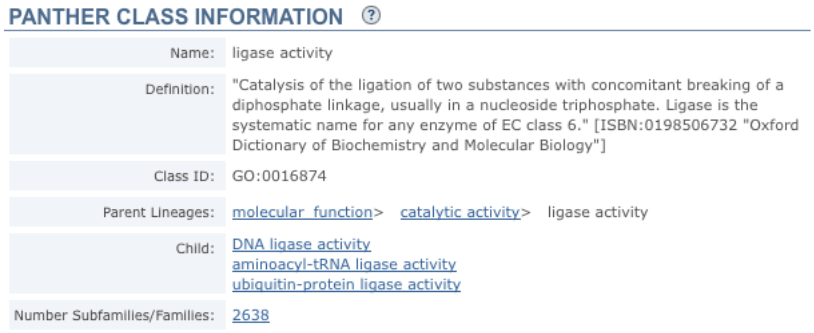

PANTHER CLASS INFORMATION
Name:
ligase activity
Definition:
"Catalysis of the ligation of two substances with concomitant breaking of a diphosphate linkage, usually in a nucleoside triphosphate. Ligase is the systematic name for any enzyme of EC class 6." [ISBN:0198506732 "Oxford Dictionary of Biochemistry and Molecular Biology"]
Class ID:
GO:0016874
Parent Lineages:
molecular_function> catalytic activity> ligase activity
Child:
DNA ligase activity
aminoacyl-tRNA ligase activity
ubiquitin-protein ligase activity
Number Subfamilies/Families:
2638

Figure 3.17: *Ontology term detail page.*

3.3.4 Pathway description page

This page provides details of the pathway. To reach this page, just click the pathway name link anywhere on the PANTHER website, and then click the "Pathway Description" tab on the page. You can toggle between this page and Pathway Diagram page.

This page provides following information about a pathway (Figure 3.18):

Definition A detailed description of the pathway.

Pathway Accession

Components The number of pathway components within the pathway. A pathway component is a group of homologous proteins across various organisms that participate in the same specific biochemical reactions within the pathway.

Clicking on the number will lead to a Pathway component list page (section 3.2.4).

Subfamilies The number of subfamilies that have at least one training sequence associated directly with the pathway. Clicking on the number will lead to a Subfamily list page (section 3.2.2).

Associated Sequences The number of training sequences that are directly associated with the pathway by manual curation. Clicking on the number will lead to the Gene list page (section 3.2.1).

References A list of literature references or websites used to generate the pathway diagram.

Author The author who generated the pathway diagram, and curated training sequences to pathway associations.

Released On The date the pathway was generated and released.

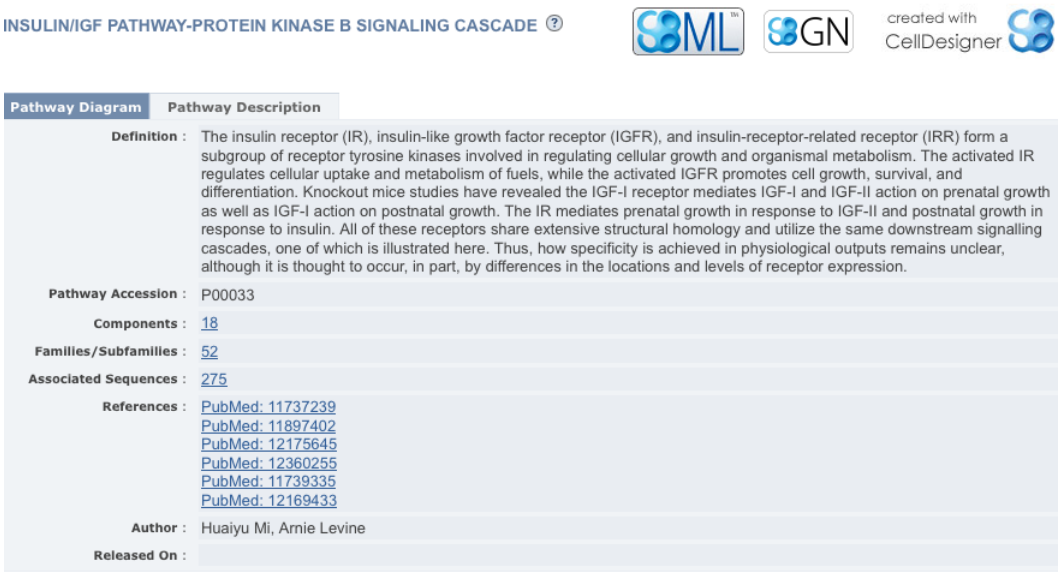

INSULIN/IGF PATHWAY-PROTEIN KINASE B SIGNALING CASCADE
created with CellDesigner
Pathway Diagram
Pathway Description
Definition : The insulin receptor (IR), insulin-like growth factor receptor (IGFR), and insulin-receptor-related receptor (IRR) form a subgroup of receptor tyrosine kinases involved in regulating cellular growth and organismal metabolism. The activated IR regulates cellular uptake and metabolism of fuels, while the activated IGFR promotes cell growth, survival, and differentiation. Knockout mice studies have revealed the IGF-I receptor mediates IGF-I and IGF-II action on prenatal growth as well as IGF-I action on postnatal growth. The IR mediates prenatal growth in response to IGF-II and postnatal growth in response to insulin. All of these receptors share extensive structural homology and utilize the same downstream signalling cascades, one of which is illustrated here. Thus, how specificity is achieved in physiological outputs remains unclear, although it is thought to occur, in part, by differences in the locations and levels of receptor expression.
Pathway Accession : P00033
Components : 18
Families/Subfamilies : 52
Associated Sequences : 275
References : PubMed: 11737239
PubMed: 11897402
PubMed: 12175645
PubMed: 12360255
PubMed: 11739335
PubMed: 12169433
Author : Huaiyu Mi, Arnie Levine
Released On :

Figure 3.18: *Pathway Description page.*

3.3.5 Pathway molecular class (component) detail page

A pathway component represents a group of homologous proteins across various organisms that participate in the same specific biochemical reactions within the pathway. A pathway component is always associated with genes within the PANTHER libraries, thus, it is linked directly or indirectly to the PANTHER family/subfamily statistical models.

This page provides details of a pathway component in two separate parts (Figure 3.19):

1. Pathway component

   Name The name that appears on the pathway diagram. It is usually an acronym or a short version of the full name.

   Full name The complete, more descriptive version of the name.

   Synonyms All other names used to describe the component.

   Definition A short description of the component.

   Category ID Each component is considered as a pathway ontology term, and is assigned with an ID. It is usually a 5 digit number preceded by letter(s) to indicate the type of category, e.g., G for DNA or RNA, and P for protein.

   Upstream A list of component(s) that modifies the current component in the pathway. It usually implies that these components have protein-protein interaction.

   Downstream A list of component(s) that is modified by the current component in the pathway. It usually implies that there are protein-protein interaction involved.

   Number Subfamilies Number of subfamilies that have training sequences directly associated with the component. Clicking on the number will lead to the subfamily list page (section 3.2.2).

   Number Associated Sequences Number of genes that are directly associated with the component. Clicking on the number will lead gene list page (section 3.2.1).

   GO classifications and PANTHER protein class The ontology terms classified to the genes that are associated to the pathway component. Clicking on an ontology terms will lead to the *Ontology Term Details page* (section 3.3.3).

   References A list of literature references or websites used to generate the pathway diagram.

2. Sequence Association A list of all training sequences that are associated with the component by manual curation. The list includes the following columns:

   PANTHER Subfamily PANTHER subfamily Ids. It links to PANTHER subfamily information page.

   Associated Sequences Training sequence ID. It links to NCBI Entrez page.

   Definition The gene definition for the training sequence

   Evidence Code* See below.

   Evidence A list papers with experimental results indicating the involvement of the sequence in the pathway. It is usually in the form of PubMed IDs.

*What is an Evidence Code?

Evidence Codes are used to denote the type of the evidence used for associating sequences to pathway components. The PANTHER Evidence Code is modeled after the GO Evidence Codes (http://www.geneontology.org/GO.evidence.shtml).

PANTHER Pathway > p53 pathway > Mdm2

PATHWAY COMPONENT DETAILS

Name:	Mdm2
Definition:	MDM2 can inactivate several functions of the tumor suppressor p53 and can degrade p53.
Category ID:	G01563
Synonyms:	
Upstream:	p53
Downstream:	Mouse double minute 2 (Mdm2)
Number Families/Subfamilies:	3
Number of Associated Sequences:	11
GO Molecular Function:	
GO Biological Process:	
GO Cellular Component:	
PANTHER protein class:	
References:	PubMed:14707284 PubMed:1614537

SEQUENCE ASSOCIATION

Panther Subfamily	Associated Sequence	Definition	Confidence Code	Evidence
PTHR10360:SF9	ENSBTAP00000013766	null	IGI	Medline:7477327
	Q00987	E3 ubiquitin-protein ligase Mdm2	IGI	Medline:7477327
	P56950	E3 ubiquitin-protein ligase Mdm2	IGI	Medline:8146175
	ENSBTAP00000013766	null	IGI	Medline:8146175
	P23804	E3 ubiquitin-protein ligase Mdm2	IGI	Medline:8146175
	Q561Z0	Mdm2 protein	IGI	Medline:8146175
PTHR10360:SF10	Q5XIN1	Protein Mdm4	ISS	
PTHR10360:SF9	Q561Z0	Mdm2 protein	IGI	Medline:7477327
	Q00987	E3 ubiquitin-protein ligase Mdm2	IGI	Medline:8146175

Figure 3.19: *Pathway Molecule Class (component) detail page.*

3.4 Graph and diagram pages

3.4.1 Pie charts

A pie chart can be generated from a gene list (section 3.2.1). Whole genome pie charts for all 48 organisms in PANTHER have been pregenerated and can be accessed from the *Genes and Orthologs* subject home page (section 3.1.1). Here is what you can do on this page (Figure 3.20).

- If it is a whole genome pie chart, you can choose the ontology you want to display from the *Select ontology* drop-down menu.
- From the *Select species* drop-down menu, you can choose one of the 48 organism you want to display.
- You can see the category names and statistics by mouse-over the pie chart selection. See detailed explanation of the statistics below.
- You can drill down to pie chart of child categories by clicking on a pie chart selection.
- You can retrieve a gene list by clicking on the chart legend link on the right side.
- The page also provides links to allow you to convert the pie chart to bar chart or to export the data as .txt file, so you can use your own program to create charts.

How to read the pie chart statistics?

When you mouse-over the pie chart selection, the category name and a series of numbers are displayed. These numbers are the statistics for the category you selected. Here is how the numbers are from.

1. The first number is the number of genes that are classified to this category. In our example in figure 3.20, it is 2067
2. The second number is the percent of genes classified to this category over the total number of genes. Right above the pie chart, it shows that the total number of genes is 17181. Remember this is the total number of human genes in the PANTHER library.
3. The third number is the percent of genes classified to this category over total number of class hits, in this case 13148. Class hit means independent ontology terms. If a gene is classified to 2 ontology terms that are not parent or child to each other, it counts as 2 class hits.

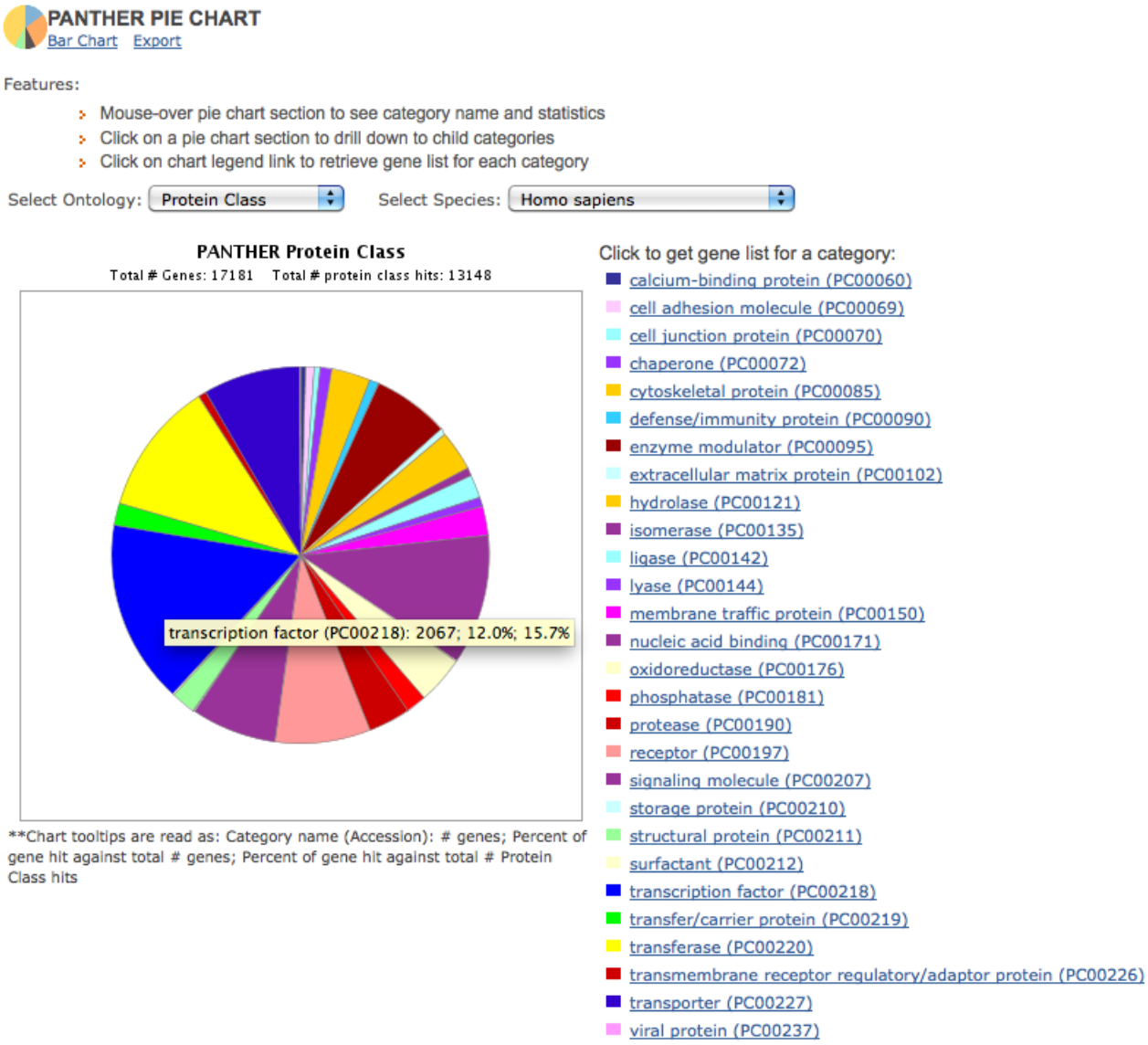

PANTHER PIE CHART
Bar Chart Export
Features:
Mouse-over pie chart section to see category name and statistics
Click on a pie chart section to drill down to child categories
Click on chart legend link to retrieve gene list for each category
Select Ontology: Protein Class
Select Species: Homo sapiens
PANTHER Protein Class
Total # Genes: 17181 Total # protein class hits: 13148
transcription factor (PC00218): 2067; 12.0%; 15.7%
**Chart tooltips are read as: Category name (Accession): # genes; Percent of gene hit against total # genes; Percent of gene hit against total # Protein Class hits
Click to get gene list for a category:
calcium-binding protein (PC00060)
cell adhesion molecule (PC00069)
cell junction protein (PC00070)
chaperone (PC00072)
cytoskeletal protein (PC00085)
defense/immunity protein (PC00090)
enzyme modulator (PC00095)
extracellular matrix protein (PC00102)
hydrolase (PC00121)
isomerase (PC00135)
ligase (PC00142)
lyase (PC00144)
membrane traffic protein (PC00150)
nucleic acid binding (PC00171)
oxidoreductase (PC00176)
phosphatase (PC00181)
protease (PC00190)
receptor (PC00197)
signaling molecule (PC00207)
storage protein (PC00210)
structural protein (PC00211)
surfactant (PC00212)
transcription factor (PC00218)
transfer/carrier protein (PC00219)
transferase (PC00220)
transmembrane receptor regulatory/adaptor protein (PC00226)
transporter (PC00227)
viral protein (PC00237)

Figure 3.20: *The pid chart of the entire human genome in PANTHER Protein Class.*

3.4.2 Pathway diagram

3.4.2.1 *PNATHER Pathway Applet Basics*

The pathway diagram on this page is interactive, and is displayed in an applet window. The window is divided into 2 panels (Figure 3.21).

- The left panel shows a list of molecules shown in the diagram. The types of the molecules are in parentheses, e.g., PROTEIN, SIMPLE_MOLECULE, GENE, etc.
- The right panel is the diagram. All objects on the diagram are interactive, so you can customize the diagram according to your preferences. The modified diagram can be saved as png file as described in section 3.4.2.2.

There are three views for the diagram.

- Interactive Standard/Activity Flow - When this tab is selected, the applet allows you to view two views that are not SBGN compliant:

– Standard View – This view faithfully reflects the original diagram generated by the CellDesigner software. It captures detailed molecular events or biochemical reactions within the pathway. It is a little different from the conventional way of illustrating pathways. For example, if a protein A activates protein B, it will be illustrated as protein A catalyzing a transition reaction of protein B from a non-active to an active form.

– Activity Flow – This view uses simpler or more conventional relationship notations to illustrate information flowchart of the pathway. It captures mainly 4 relationships between 2 proteins: activation, inhibition, neutral, and unknown. It is derived based on the Standard View. In cases where a reaction can not be simply illustrated as one of the above four, it will keep it in the original Standard view. This view is not compliant with SBGN Activity Flow Diagram.

The Activity Flow view is the default view of the pathways. You can toggle the two view by clicking the tab on the top of the applet window.

- SBGN Image View - When this tab is selected, a static (non-interactive) image of SBGN Process Description Diagram is displayed.

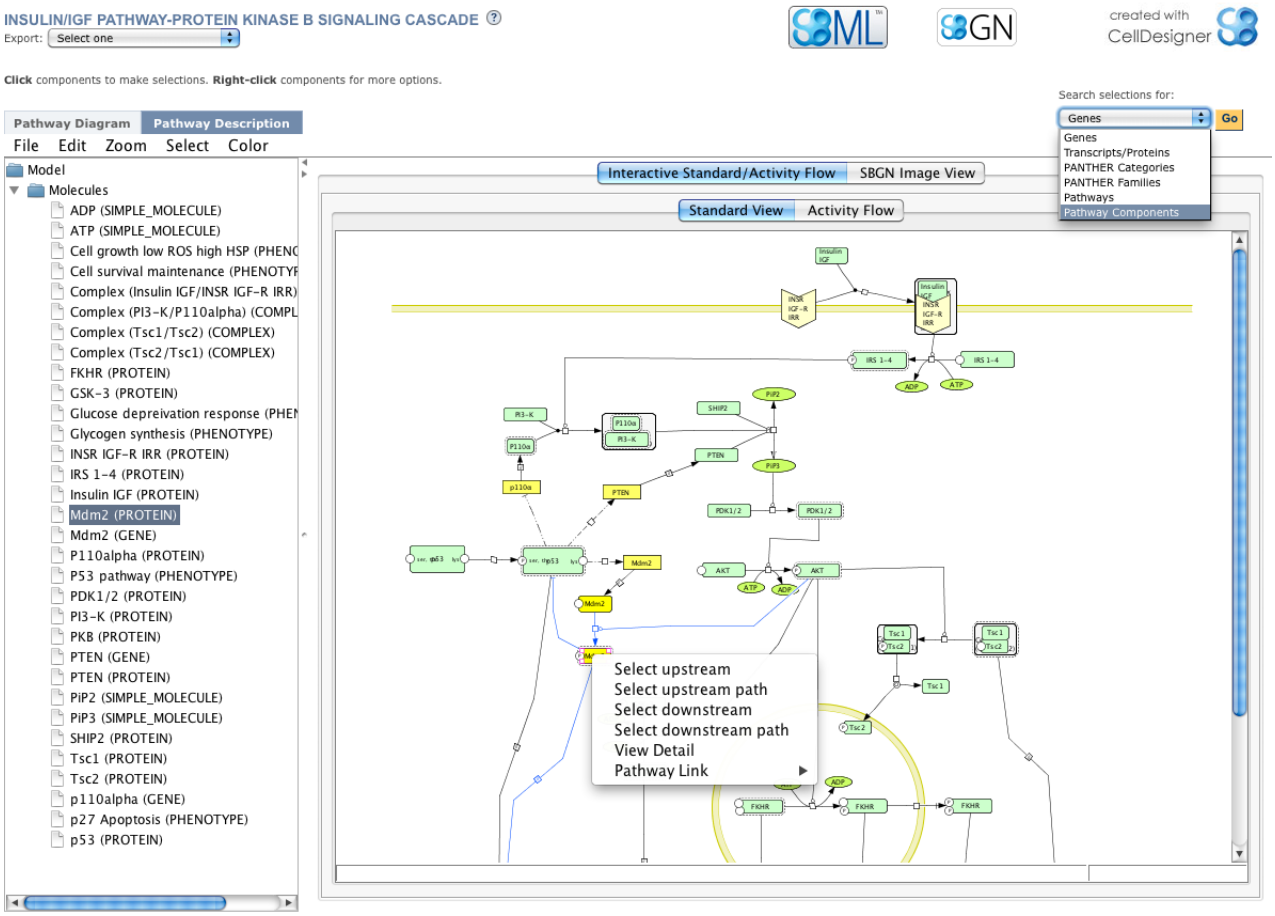


Figure 3.21: *Pathway Diagram page.*

3.4.2.2 Export data from this page

Users can export pathway related data using the following options.

- From the *Export* drop-down menu, all the pre-generated images and SBML file of the pathway can be exported
- From the applet panel, under "File", select "Export image" to export the current pathway diagram. This function allows you to export the modified image.

3.4.2.3 Using the pathway diagram

The diagram page allows you to make selections on pathway components, and then generate gene lists, transcript lists, or family lists.
You can use any of the following ways to select a single or multiple components.

1. You can select a component by clicking on the name of the component in the left (list) panel of the applet. If you want to select more than one component, press and hold the control key while you select. You can clear your selection by going to "Select -¿ Clear all" on the applet menu bar.
2. You can select components from the right diagram panel by clicking on the component symbol. You can clear your selection by clicking on the selected item again. You don't need to hold the control key for multiple selections.
3. Under "Select" applet menu bar, you can do the following.
   - Select all proteins, genes and RNA. Since only proteins, genes and RNA in PANTHER pathway are associated with PANTHER family genes, you have an option to select all of them by one click.
   - Select all. This function allows you to select all items, including simple molecules, ions, phenotypes, etc.
   - Clear all. You can use this function to clear all your selections.
4. By right-clicking on the component, you can make the following selections.
   - Upstream or Downstream component. This function will select the immediate or downstream component. If the selected component appears in multiple places in the same pathway, they will be highlighted also.
   - Upstream or Downstream path. This function will select all components in the entire upstream or downstream path.

Once the selections are made, you can generate any of the following lists by making the selection to the *Search selection for* drop-down menu on the right side of the page.

- Genes: A list if genes whose proteins hit PANTHER family and subfamily models with at least one training sequence associated with the selected pathway component.
- Transcripts/proteins: A list of proteins that hit PANTHER family and subfamily models with at least one training sequence associated with the selected pathway component.
- PANTHER Ontology Terms: A list of PANTHER Molecular Function and Biological Process ontology terms that are classified to PANTHER family or subfamily models with at least one training sequence associated with the selected pathway component.
- PANTHER Families: A list of PANTHER families or subfamilies with at least one training sequence associated with the selected pathway component.
- Pathway
- Pathway Components

3.4.3 TreeViewer

PANTHER trees allow you to explore the relationships between sequences in a particular family, as well as view some of the key information that was used to annotate the families and subfamilies. The current PANTHER release (version 7.0) contain publicly-available protein sequences from UniProt, Refseq, ENSEMBL and Model Organism Databases.

The Tree Viewer has two panels that are mapped to each other (Figure 3.22). The left panel displays the relationship between the different sequences, and is labeled as *Tree*. The right panel has two views. The first is the "attribute table", which is labeled as *Grid* and contains information for the each corresponding sequence in the tree panel. The second view is the "multiple sequence alignment" view, which is labeled as *MSA* (Figure 3.24). You can click the the button to toggle between the two views.

3.4.3.1 *Trees*

In the tree panel on the left of the Tree Viewer, you can explore that phylogenetic relationships among different genes within the family. The longer the (horizontal) branch length, the more distant the groups joined by those branches. Vertical branch length is fixed for ease of viewing together with the information on the panel.
There are 4 types of nodes in the phylogenetic tree:

- Blue diamonds - Subfamily nodes. Subfamilies are usually colored to help distinguish between different subfamilies. Aside from this, the subfamily color does not have any special significance.
- Orange circles - Gene duplication nodes.
- Green circles - Speciation nodes.
- End of the horizontal branch - Leaf nodes. They usually point to the sequence IDs

The following are a few things that you can do to these nodes:

1. When mouse-over these nodes, the subfamily annotation information is displayed. For the the subfamily node (blue diamond), it displays the subfamily id and name. For all other node types, it displays which subfamily it belongs to.
2. When a node is clicked (except for the leaf node), it collapses.
    - A collapsed subfamily view provides a high-level view of the tree, in which subfamilies are the most specific "leaves" of the tree. The subfamily name given by curators appears in the "sf_name" column of the Collapsed view. The GO and PANTHER protein class terms are shown in the appropriate columns also.
    - A collapsed non-subfamily node is shown as a red circle. The annotation node (AN) id is shown in the "Protein id" column, and the common GO and protein class annotation for all sequence below the collapsed node is shown in the appropriate columns.

Under the "Tree" menu, you can do the followings (Figure 3.22):

- Collapse lowest level of Subfamilies - A PANTHER subfamily node can have other subfamily node(s) as descendants. This option collapses all subfamily nodes that do not contain another subfamily node. All subfamily nodes that nest another subfamily node remain expanded.
- Expand all Subfamily nodes - All subfamily nodes are expanded. All non-subfamily nodes remain unchanged.

- Expand all Nodes - All nodes, including non-subfamily nodes can be expanded
- Use distances - If this option is selected (a check mark shows in front of the option), the tree branches are shown based on the distances calculated in tree building algorithm (GIGA) [5]. If this option is unchecked, the tree branch will be shown with fixed length.
- Ladder Top - Re-arrange the tree so that the node with more descendants are always above the nodes with less.
- Ladder Bottom - Opposite to the above. Re-arrange the tree so that the node with more descendants are always below the nodes with less.
- Latter Original - Set the order of the tree to what was loaded from the server. It is different from either Ladder Top or Ladder Bottom.
- Colore Subfamilies

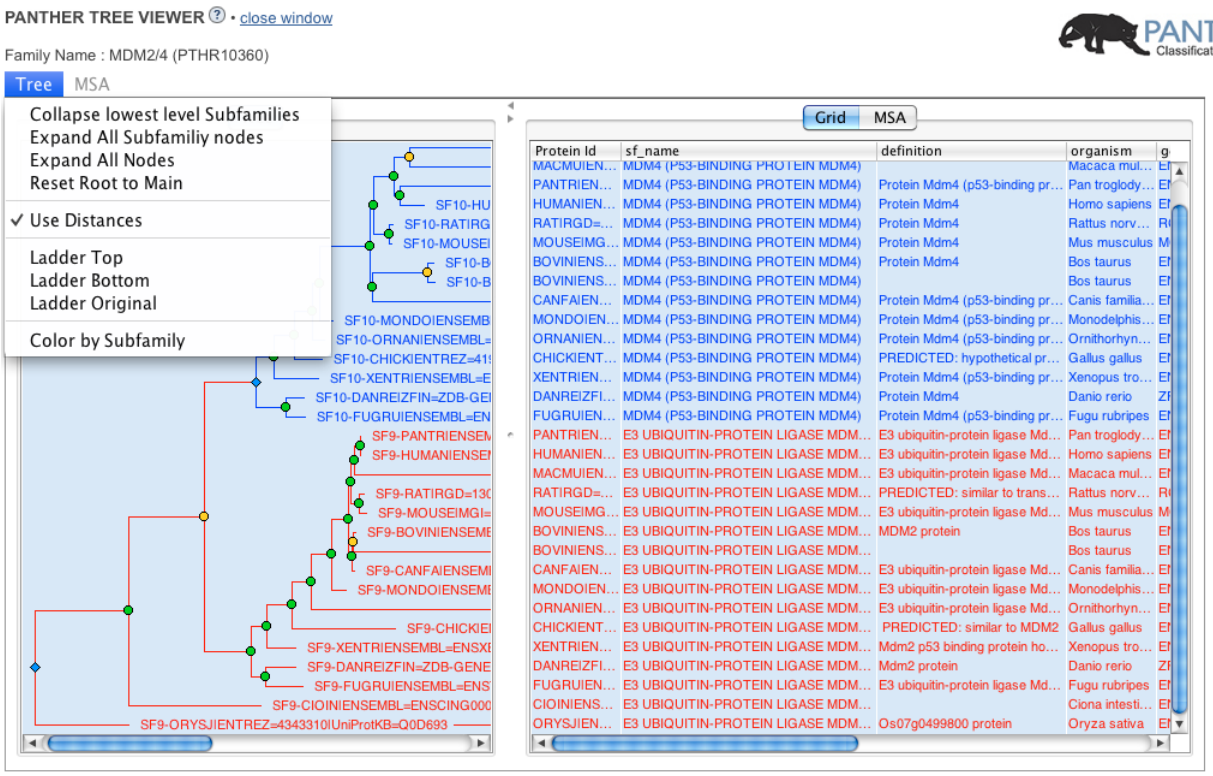


Figure 3.22: *A phylogenetic tree of PANTHER family PTHR10360 viewed in the tree viewer.*

3.4.3.2 Attribute table

The attribute table contains one row for each sequence in the tree. Each column displays a different attribute of the sequences as follows (Figure 3.23).:

- Protein id - This is the long id for each sequence. The format of the id is as below: organism|gene DB=gene id|protein DB=protein id. When the tree is collapsed, this column shows the annotation node (AN) ids.
- sf_name - The name annotated for the subfamily.

- definition - This is the brief definition line parsed out from the protein database where the protein sequence is from.
- organism - This is the organism from which the sequence was derived. Click on the organism name to open the full taxonomy record for that organism.
- gene id
- gene symbol
- OrthoMCL - The OrthoMCL id that the gene is predicted to belong to.
- Molecular function, Biological Process and Cellular component - These are Gene Ontology terms from PANTHER GO Slim describing the function of the gene product.
- Protein Class - This is a PANTHER Index term describing protein classes.

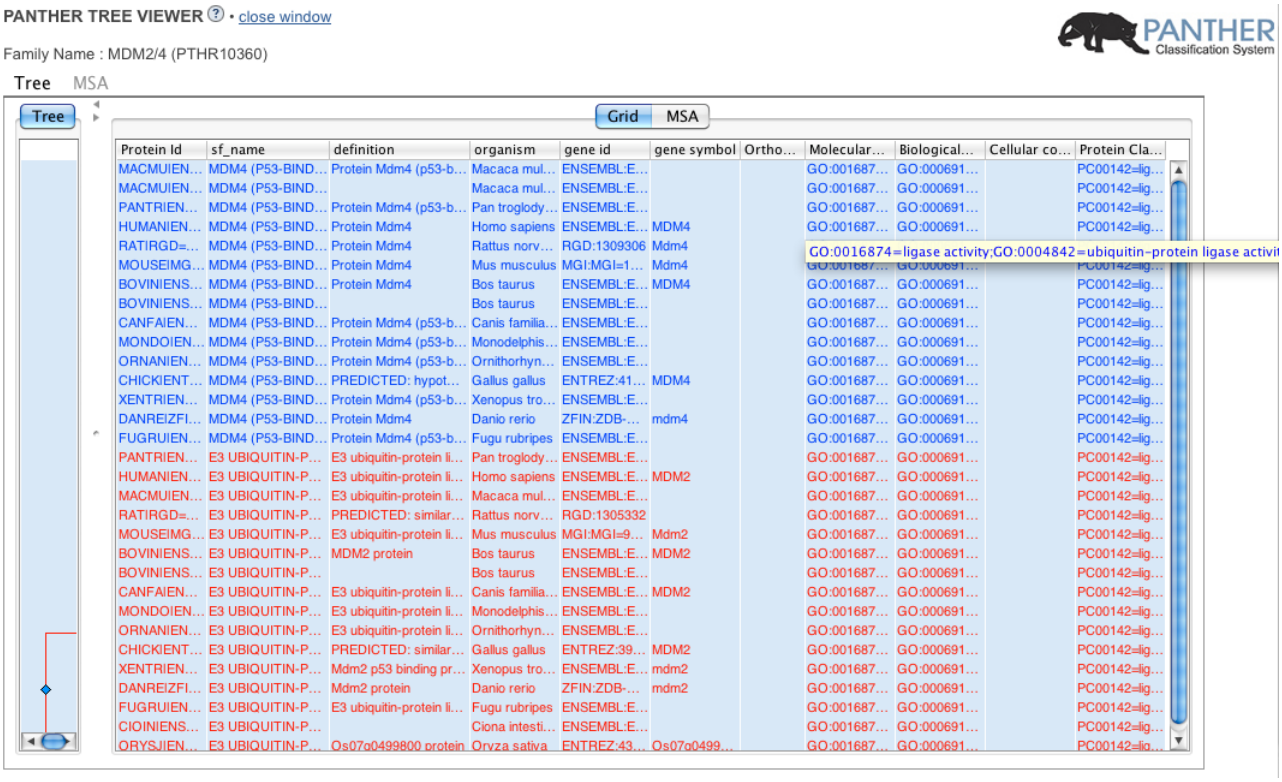

PANTHER TREE VIEWER • close window
Family Name : MDM2/4 (PTHR10360)
Tree MSA
PANTHER Classification System
Tree
Grid MSA
Protein Id | sf_name | definition | organism | gene id | gene symbol | Ortho... | Molecular... | Biological... | Cellular co... | Protein Cla...
GO:0016874=ligase activity;GO:0004842=ubiquitin-protein ligase activi

Figure 3.23: *A tree viewer with expanded attribute table.*

3.4.3.3 MSA

When the "MSA" button is clicked on the right panel, the multiple sequence alignment view is shown. You can toggle back to the attribute table view by clicking the "Grid" button.

PANTHER multiple sequence alignments (MSAs) are the basis for the PANTHER distance trees, and therefore of the family/subfamily classification. MSA is built using the MAFFT software [13]. MSA is aligned across the entire length of all sequences in the family. The evolutionarily conserved part of the alignment is used to build HMM, and it is called the *match state*, which is designated with *uppercase letters*. The other less conserved region is designated by *lowercase letters*, and is called *non-match state*. If a sequence misses a position in the match

state, it is called a *delete state* and is designated by a *dash*. If a sequence misses a position in the non-match state, it is called an *insert state* and is designated by a *dot*.

In the MSA menu of the MSA view, you can make the following selections.

- Percentage Identity for Entire Alignment - This is the default selection and shows the alignment across the entire length of the all sequences.
- Percentage Identity for Match States - This only shows the alignment for the match states

The MSA uses uppercase letters to indicate match state positions, lowercase letters to indicate non-match state positions, dots to indicate insert state positions, and dashes to indicate delete state positions.

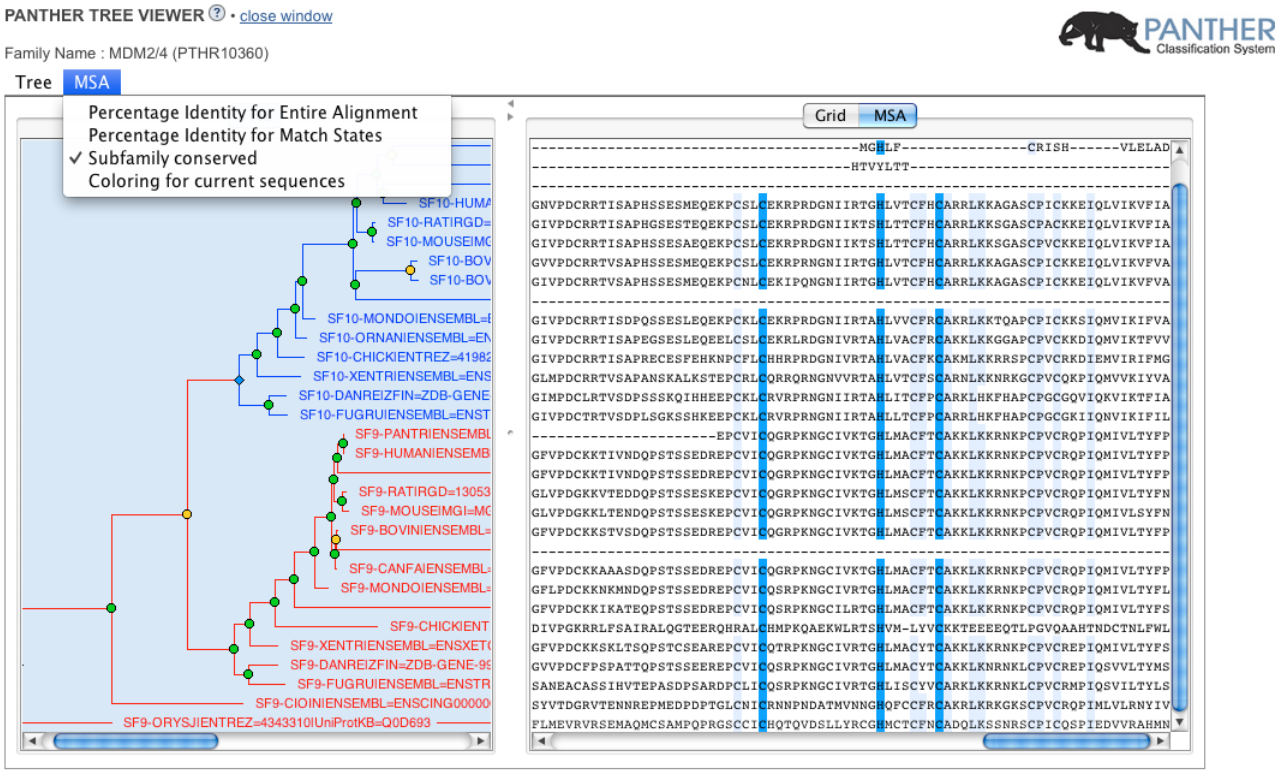


Figure 3.24: *The multiple sequence alignment of PANTHER family PTHR10360 viewed in the tree viewer.*

Chapter 4

PANTHER Basics

4.1 Download

This is the PANTHER FTP site. You can reach this page by clicking the "Downloads" link on the *Quick links* panel on the left side of the Home page. You can download the following PANTHER data and tools.

- PANTHER HMM library - The current version is 7.0 that contains the PANTHER HMMs for 6594 protein family HMMs, divided into 62,972 functionally distinct protein subfamily HMMs. The PANTHER HMMs are in the HMMER format. There is also a version of the library available that can be used for the Coding SNP Analysis tool.
- PANTHER HMM scoring tool - Download the PANTHER scoring tool, to score protein sequences against the entire PANTHER HMM library and analyze your sequences. The HMM library should be downloaded separately.
- PANTHER cSNP scoring tool - Download the PANTHER Coding SNP Analysis tool to estimate the likelihood that a particular nonsynonymous (amino-acid changing) coding SNP will cause a functional impact to the protein. A special version of HMM library for this tool should be downloaded separate.
- GIGA - GIGA is an efficient tree building program that allows phylogenetic reconstruction of very large gene families and determination of orthologs on a large scale. It is used to generate trees in the PANTHER Classification System.
- PANTHER Pathway - Contains 165 regulatory and metabolic pathways, each with subfamilies and protein sequences mapped to individual pathway components. All pathway diagrams are abailable in both SBML and BioPAX level 3 format. SBML files, BioPAX files and Protein Sequence Association data are available for download.
- PANTHER Sequence Classification file - Contains the PANTHER family, subfamily, GO slim, PANTHER protein class, and pathway classifications for the the complete genomes derived from the 48 organisms used in the PANTER database.
- PANTHER HMM Classification file - Contains the PANTHER family/subfamily name, and the GO slim, PANTHER protein class, and pathway classifications for every PANTHER protein family and subfamily in PANTHER HMM library.
- PAINT - PAINT is a java application for viewing tree files. Minimum Java version 1.4 required. To install, uncompress the download file and execute launchTAV.bat for windows or sh launchTAV.sh for MAC.

4.2 Workspace

If you have registered and have a *Workspace* account in PANTHER, here is a description about how you can use it.

You can export any list from the list page to the *Workspace*. Typically, after a keyword search, prowler search, or batch search, you would like to save the resulting list so that you do not have to perform such search over and over. *Workspace* allows you to do so. When you use the export to workspace function on any of the list page (see description is section 3.2.1), you will be asked to provide a name (required) and a description (optional) to the list.

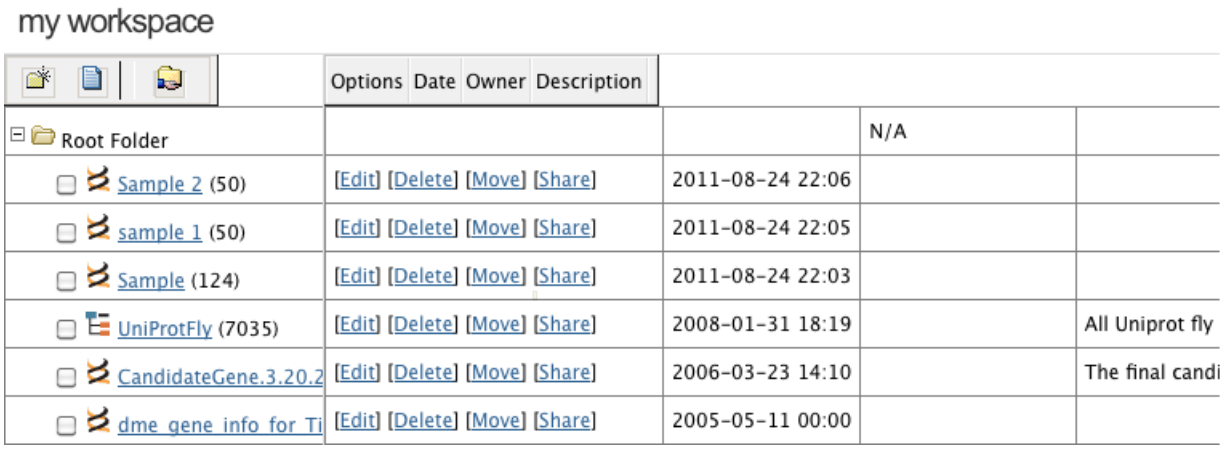


Figure 4.1: *Screenshot of a sample workspace page.*

To access your *Workspace*, you just need to click the "Workspace" link at the PANTHER Home page (Section]refsec:ov) and login. You will then reach the main Workspace page (Figure 4.1).

The exported list is in the "root folder" by default. You can organize the lists by using the "Move" link in the second column of the table to move the list to the appropriate folder. You can also create a new folder by clicking the "create new folder" icon on the top left part of the table.

Here are some other things you can do to your lists.

- Go to the list by clicking the list name in the first column. The number in the parenthesis indicates the number of entries (e.g., genes) in the list.
- Share the list with other registered PANTHER users by clicking the *Share* link or the icon. You then need to enter the username of the *Workspace* that you would like to share.
- Modify the name and description by clicking the *Edit* button.
- Delete the list.

Appendix A

Systems requirements

- PC Users (Recommended):
  - Windows 2000, or Windows XP
  - Microsoft Internet Explorer 6.x
  - Java version 1.4.2 is required
- Macintosh Users:
  - A G4 Macinosh with MacOS 10.3.7 is required
  - Safari version 1.2.4
  - Java version 1.4.2 is required
- For both:
  - Minimum of 128 MB RAM, 256MB RAM recommended
  - JavaScript, Java applets and cookies must be enabled in your browser
  - Java applet runtime parameters set to -ms128m -mx512m -Xss16m
  - Screen resolution of at least 800x600 dpi is strongly recommended

Bibliography

[1] Huaiyu Mi, Anushya Muruganujan, and Paul D. Thomas. Panther in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Research*, 41(D1):D377–D386, 1 2013.

[2] Huaiyu Mi, Anushya Muruganujan, John T. Casagrande, and Paul D. Thomas. Large-scale gene function analysis with the panther classification system. *Nat Protoc*, 8(8):1551–66, 8 2013.

[3] Huaiyu Mi and Paul Thomas. Panther pathway: an ontology-based pathway database coupled with data analysis tools. *Methods Mol Biol*, 563:123–40, 2009.

[4] Paul D. Thomas, Anish Kejariwal, Nan Guo, Huaiyu Mi, Michael J. Campbell, Anushya Muruganujan, and Betty Lazareva-Ulitsky. Applications for protein sequence-function evolution data: mrna/protein expression analysis and coding snp scoring tools. *Nucleic Acids Res*, 34(Web Server issue):W645–50, 7 2006.

[5] Paul D. Thomas. Giga: a simple, efficient algorithm for gene tree inference in the genomic age. *BMC Bioinformatics*, 11:312, 2010.

[6] J. A. Blake, M. Dolan, H. Drabkin, D. P. Hill, Ni Li, D. Sitnikov, S. Bridges, S. Burgess, T. Buza, F. McCarthy, D. Peddinti, L. Pillai, S. Carbon, H. Dietze, A. Ireland, S. E. Lewis, C. J. Mungall, P. Gaudet, R. L. Chrisholm, P. Fey, W. A. Kibbe, S. Basu, D. A. Siegele, B. K. McIntosh, D. P. Renfro, A. E. Zweifel, J. C. Hu, N. H. Brown, S. Tweedie, Y. Alam-Faruque, R. Apweiler, A. Auchinchloss, K. Axelsen, B. Bely, M. C Blatter, C. Bonilla, L. Bouguerleret, E. Boutet, L. Breuza, A. Bridge, W. M. Chan, G. Chavali, E. Coudert, E. Dimmer, A. Estreicher, L. Famiglietti, M. Feuermann, A. Gos, N. Gruaz-Gumowski, R. Hieta, C. Hinz, C. Hulo, R. Huntley, J. James, F. Jungo, G. Keller, K. Laiho, D. Legge, P. Lemercier, D. Lieberherr, M. Magrane, M. J. Martin, P. Masson, P. Mutowo-Muellenet, C. O'Donovan, I. Pedruzzi, K. Pichler, D. Poggioli, P. Porras Millán, S. Poux, C. Rivoire, B. Roechert, T. Sawford, M. Schneider, A. Stutz, S. Sundaram, M. Tognolli, I. Xenarios, R. Foulgar, J. Lomax, P. Roncaglia, V. K. Khodiyar, R. C. Lovering, P. J. Talmud, M. Chibucos, M. Gwinn Giglio, H. Y Chang, S. Hunter, C. McAnulla, A. Mitchell, A. Sangrador, R. Stephan, M. A. Harris, S. G. Oliver, K. Rutherford, V. Wood, J. Bahler, A. Lock, P. J. Kersey, D. M. McDowall, D. M. Staines, M. Dwinell, M. Shimoyama, S. Laulederkind, T. Hayman, S. J Wang, V. Petri, T. Lowry, P. D'Eustachio, L. Matthews, R. Balakrishnan, G. Binkley, J. M. Cherry, M. C. Costanzo, S. S. Dwight, S. R. Engel, D. G. Fisk, B. C. Hitz, E. L. Hong, K. Karra, S. R. Miyasato, R. S. Nash, J. Park, M. S. Skrzypek, S. Weng, E. D. Wong, T. Z. Berardini, E. Huala, H. Mi, P. D. Thomas, J. Chan, R. Kishore, P. Sternberg, K. Van Auken, D. Howe, and M. Westerfield. Gene ontology annotations and resources. *Nucleic Acids Res*, 41(Database issue):D530–5, 1 2013.

[7] Raymond J. Cho and Michael J. Campbell. Transcription, genomes, function. *Trends in Genetics*, 16(9):409 – 415, 2000.

[8] A. G. Clark, S. Glanowski, R. Nielsen, P. D. Thomas, A. Kejariwal, M. A. Todd, D. M. Tanenbaum, D. Civello, F. Lu, B. Murphy, S. Ferriera, G. Wang, X. Zheng, T. J. White, J. J. Sninsky, M. D. Adams, and M. Cargill. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science*, 302(5652):1960–3, 12 2003.

[9] Vamsi K. Mootha, Cecilia M. Lindgren, Karl-Fredrik F. Eriksson, Aravind Subramanian, Smita Sihag, Joseph Lehar, Pere Puigserver, Emma Carlsson, Martin Ridderstråle, Esa Laurila, Nicholas Houstis, Mark J. Daly, Nick Patterson, Jill P. Mesirov, Todd R. Golub, Pablo Tamayo, Bruce Spiegelman, Eric S. Lander, Joel N. Hirschhorn, David Altshuler, and Leif C. Groop. Pgc-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet*, 34(3):267–73, 7 2003.

[10] L. R. Brunham, R. R. Singaraja, T. D. Pape, A. Kejariwal, P. D. Thomas, and M. R. Hayden. Accurate prediction of the functional significance of single nucleotide polymorphisms and mutations in the abca1 gene. *PLoS Genet*, 1(6):e83, 12 2005.

[11] Paul D. Thomas, Michael J. Campbell, Anish Kejariwal, Huaiyu Mi, Brian Karlak, Robin Daverman, Karen Diemer, Anushya Muruganujan, and Apurva Narechania. Panther: a library of protein families and subfamilies indexed by function. *Genome Res*, 13(9):2129–41, 9 2003.

[12] Paul D. Thomas and Anish Kejariwal. Coding single-nucleotide polymorphisms associated with complex vs. mendelian disease: evolutionary evidence for differences in molecular effects. *Proc Natl Acad Sci U S A*, 101(43):15398–403, 10 2004.

[13] K. Katoh, K. Kuma, H. Toh, and T. Miyata. Mafft version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res*, 33(2):511–8, 2005.

Index

Activity Flow, 50
Alternative ID, 40
annotation data, 5
Applet, *see* PANTHER Pathway Applet
attribute table, 52, 53

batch ID search, 8, 30
binomial distribution test, 8, 12
biological process, 20, 40, 42
BioPAX, 32
Bonferroni correction, 8
Bonferroni test, 14
branch length, 52
browse, 19

cellular component, 20, 40, 42
citing PANTHER, 1, 3
closely related, 23
collapse, 52
Component, *see* pathway molecule class
component, *see* pathway molecule class, 37, 45, 46
Component Accession, 37
Component Name, 38
cSNP scoring, 24

delete state, 55
distantly related, 23
Download, 3, 56
downstream, 38, 46, 51
downstream path, 51

Evidence, 47
Evidence Code, 47
expand, 52, 53
expected value, 12
export, 35, 50

Family ID, 35
family list page, 35
Family Name, 35
family/subfamily detail page, 42
feedback email, 2
file format, 6, 7, 14
FTP, *see* Download

gene detail page, 40
gene duplication, 52
gene home page, 30
Gene ID, 33, 40
Gene List Analysis Tools, 4
Gene List Analysis tools
    default setting, 8, 15
    enrichment test, 14
    Functional classification, 8
    setting, 8, 12, 15, 16
gene list page, 33
Gene Name, 33, 40
Gene Ontology, 1, 39, 40, 42
Gene Symbol, 33, 40
GO, 5, 6
GO slim, 20, 39, 42, 54
GO-slim, 5

heat map, 18
hidden Markov model, 31, 42
HMM, 23, 31, 42
    length, 42
    scoring, 23, 31
home page, 2

insert state, 55
Interpro, 36

keyword search, 3, 26, 30

Ladder bottom, 53
Ladder original, 53
Ladder top, 53
LDO, *see* least diverge ortholog
leaf node, 52
least diverged ortholog, 41
LOD, *see* least diverge ortholog
login, *see* registration

Mann-Whitney test, 17
match state, 54, 55
Microsoft Excel format, 7
molecular function, 20, 40, 42
MSA, 25, 52, 54
multiple sequence alignment, 25, 52, 54

ND, 41
NIC, *see* number of independent counts
number of independent counts, 26

ontology, 20, 21
ontology home page, 33
ortholog, 41
OrthoMCL, 54
over-representation, 12, 13
overrepresentation test, 8

p-value, 13
PANTHER classification, 40
PANTHER Data, 2
PANTHER Generic Mapping File, 7
PANTHER hit, 23
PANTHER Pathway, 5
PANTHER Pathway Applet, 49
PANTHER Protein Class, 5
PANTHER protein class, 33, 36
PANTHER score, 23
PANTHER tool, 3
paralog, 41
pathway, 20, 32, 40, 42
Pathway Accession, 36, 45
pathway detail page, 45
pathway diagram, 18, 32, 49
pathway diagrams, 13
pathway home page, 32
pathway list page, 36
pathway molecule class, 32
pathway molecule class detail page, 46
Pathway Name, 37
Pdeleterious, 25
phylogenetic tree, 31, 52
pie chart, 3, 8, 30, 35, 48, 49
Process Description, 50
protein class, 20, 40
Protein ID, 40
prowler, 19, 20
PubMed, 47

rank-sum test, 14
reference, 45, 46
reference gene list, 8, 9
registration, 3

SBGN, 50
SBML, 32
sequence search, 23
simple text format, 7
sort, 34
speciation, 52
species, 20
Standard View, 50
subfamily node, 52
subPSEC, 24, 25
Supported ID, 6
synonym, 46

tool, 33
training sequence, 37, 45
tree home page, 31
tree viewer, 52

under-representation, 12, 13
upstream, 38, 46, 51
upstream path, 51

Wilcoxon Rank-Sum test, 17
workspace, 3, 33, 35, 57
    organize, 57
    share, 57